Foundations of Data Science

Final Project

Jenna Tuffnell

Due: May 13, 2025

Diabetes

```
knitr::opts_chunk$set(fig.align="center", warning=FALSE, message=FALSE)
```

Source: https://www.kaggle.com/datasets/marshalpatel3558/diabetes-prediction-dataset

The dataset we picked is on diabetes and different measurements of health such as BMI, family history of diabetes, cholesterol (LDL & HDL), age, gender, and more. The observations are of different health indicators that are all relevant to diagnosing people with diabetes, however there is no observation actually diagnosing anyone or stating who does and does not have diabetes explicitly. Instead, we decided to use HbA1c, a measure of blood glucose in the last 3-2 months, to "diagnose" people as diabetic, pre-diabetic, and non-diabetic. Originally, we wanted to use k-nearest neighbors classification to see if it correctly classifies observations as diabetic, pre-diabetic, or normal. Before doing so we used PCA to try and see if any variables were significant, and when this was unsuccessful we tested the correlation between variables and found that to be inconclusive as well. Therefore, we picked BMI, Cholesterol_HDL/LDL, waist circumference, age, and family history to be our key variables. Finally, we used k-means clustering to identify the type 1 and type 2 diabetics. This was not entirely conclusive, however, there were two distinct clusters in the end that we were able to identify as type 1 diabetic and type 2 diabetic.

R Markdown

First, we used principal component analysis to test which variables we should use for the kNN clustering.

Principal Component Analysis

```
diabetes_data <- read.csv("./Dataset/diabetes_dataset.csv", header=TRUE)
str(diabetes_data)
```

```
## 'data.frame':    10000 obs. of  21 variables:
##  $ X                          : int  0 1 2 3 4 5 6 7 8 9 ...
##  $ Age                        : int  58 48 34 62 27 40 58 38 42 30 ...
##  $ Sex                        : chr  "Female" "Male" "Female" "Male" ...
##  $ Ethnicity                  : chr  "White" "Asian" "Black" "Asian" ...
```

```
##  $ BMI                          : num  35.8 24.1 25 32.7 33.5 33.6 33.2 26.9 27 24 ...
##  $ Waist_Circumference          : num  83.4 71.4 113.8 100.4 110.8 ...
##  $ Fasting_Blood_Glucose        : num  124 184 142 167 146 ...
##  $ HbA1c                        : num  10.9 12.8 14.5 8.8 7.1 13.5 13.3 10.9 7 14 ...
##  $ Blood_Pressure_Systolic      : int  152 103 179 176 122 170 131 121 132 146 ...
##  $ Blood_Pressure_Diastolic     : int  114 91 104 118 97 90 80 83 118 83 ...
##  $ Cholesterol_Total            : num  198 262 261 183 203 ...
##  $ Cholesterol_HDL              : num  50.2 62 32.1 41.1 53.9 44.5 77.9 69.7 73.2 53.3 ...
##  $ Cholesterol_LDL              : num  99.2 146.4 164.1 84 92.8 ...
##  $ GGT                          : num  37.5 88.5 56.2 34.4 81.9 77.5 52.1 72 76.4 14.5 ...
##  $ Serum_Urate                  : num  7.2 6.1 6.9 5.4 7.4 6.4 4.7 5.6 6.2 6.9 ...
##  $ Physical_Activity_Level      : chr  "Moderate" "Moderate" "Low" "Low" ...
##  $ Dietary_Intake_Calories      : int  1538 2653 1684 3796 3161 3460 3107 2390 3844 2230 ...
##  $ Alcohol_Consumption          : chr  "Moderate" "Moderate" "Heavy" "Moderate" ...
##  $ Smoking_Status               : chr  "Never" "Current" "Former" "Never" ...
##  $ Family_History_of_Diabetes   : int  0 0 1 1 0 1 0 0 1 1 ...
##  $ Previous_Gestational_Diabetes: int  1 1 0 0 0 1 0 1 0 0 ...
```

Since all the data is not numeric, I am checking for null values and filtering out columns to only use numerical data.

```
colSums(is.na(diabetes_data))
```

```
##                              X                            Age
##                              0                              0
##                            Sex                      Ethnicity
##                              0                              0
##                            BMI            Waist_Circumference
##                              0                              0
##          Fasting_Blood_Glucose                          HbA1c
##                              0                              0
##        Blood_Pressure_Systolic       Blood_Pressure_Diastolic
##                              0                              0
##              Cholesterol_Total                Cholesterol_HDL
##                              0                              0
##                Cholesterol_LDL                            GGT
##                              0                              0
##                    Serum_Urate        Physical_Activity_Level
##                              0                              0
##        Dietary_Intake_Calories            Alcohol_Consumption
##                              0                              0
##                 Smoking_Status     Family_History_of_Diabetes
##                              0                              0
## Previous_Gestational_Diabetes
##                              0
```

```
diabetes_data2 <- select(diabetes_data, -X, -Sex, -Ethnicity, -Physical_Activity_Level, -Smoking_Status

numerical_data <- diabetes_data2[]

head(numerical_data)
```

```
##   Age  BMI Waist_Circumference Fasting_Blood_Glucose HbA1c
```

```
## 1  58 35.8                83.4                 123.9  10.9
## 2  48 24.1                71.4                 183.7  12.8
## 3  34 25.0               113.8                 142.0  14.5
## 4  62 32.7               100.4                 167.4   8.8
## 5  27 33.5               110.8                 146.4   7.1
## 6  40 33.6                96.1                  75.0  13.5
##   Blood_Pressure_Systolic Blood_Pressure_Diastolic Cholesterol_Total
## 1                     152                      114             197.8
## 2                     103                       91             261.6
## 3                     179                      104             261.0
## 4                     176                      118             183.4
## 5                     122                       97             203.2
## 6                     170                       90             152.3
##   Cholesterol_HDL Cholesterol_LDL  GGT Serum_Urate Dietary_Intake_Calories
## 1            50.2            99.2 37.5         7.2                    1538
## 2            62.0           146.4 88.5         6.1                    2653
## 3            32.1           164.1 56.2         6.9                    1684
## 4            41.1            84.0 34.4         5.4                    3796
## 5            53.9            92.8 81.9         7.4                    3161
## 6            44.5           190.0 77.5         6.4                    3460
##   Family_History_of_Diabetes Previous_Gestational_Diabetes
## 1                          0                             1
## 2                          0                             1
## 3                          1                             0
## 4                          1                             0
## 5                          0                             0
## 6                          1                             1
```

As you can see, we are now left with 14 numeric variables that we can conduct PCA with. Here we rank variables to use based on cumulative proportion, the amount of variance each component explains in the data:

```r
scaled_data <- scale(numerical_data)
pca_result <- prcomp(scaled_data, center = TRUE, scale. = TRUE)
print(summary(pca_result))
```

```
## Importance of components:
##                           PC1     PC2     PC3     PC4     PC5    PC6     PC7
## Standard deviation     1.03118 1.02629 1.01813 1.01530 1.00952 1.0048 1.00354
## Proportion of Variance 0.07089 0.07022 0.06911 0.06872 0.06794 0.0673 0.06714
## Cumulative Proportion  0.07089 0.14111 0.21021 0.27894 0.34688 0.4142 0.48132
##                           PC8     PC9   PC10    PC11    PC12    PC13    PC14
## Standard deviation     1.0017 0.99669 0.9912 0.99015 0.98718 0.98138 0.97329
## Proportion of Variance 0.0669 0.06623 0.0655 0.06536 0.06497 0.06421 0.06315
## Cumulative Proportion  0.5482 0.61444 0.6799 0.74530 0.81027 0.87448 0.93763
##                           PC15
## Standard deviation     0.96723
## Proportion of Variance 0.06237
## Cumulative Proportion  1.00000
```

```r
explained_variance <- summary(pca_result)$importance[2,]
cumulative_variance <- cumsum(explained_variance)
num_components <- which(cumulative_variance >= 0.90)[1]
print(num_components)
```

```
## PC14
##   14
```

This was inconclusive, as all components (or variables) accounted for almost exactly 7% of the data. So, to help us visualize, we also constructed a correlation plot for each variable to see if any variables seem correlated.

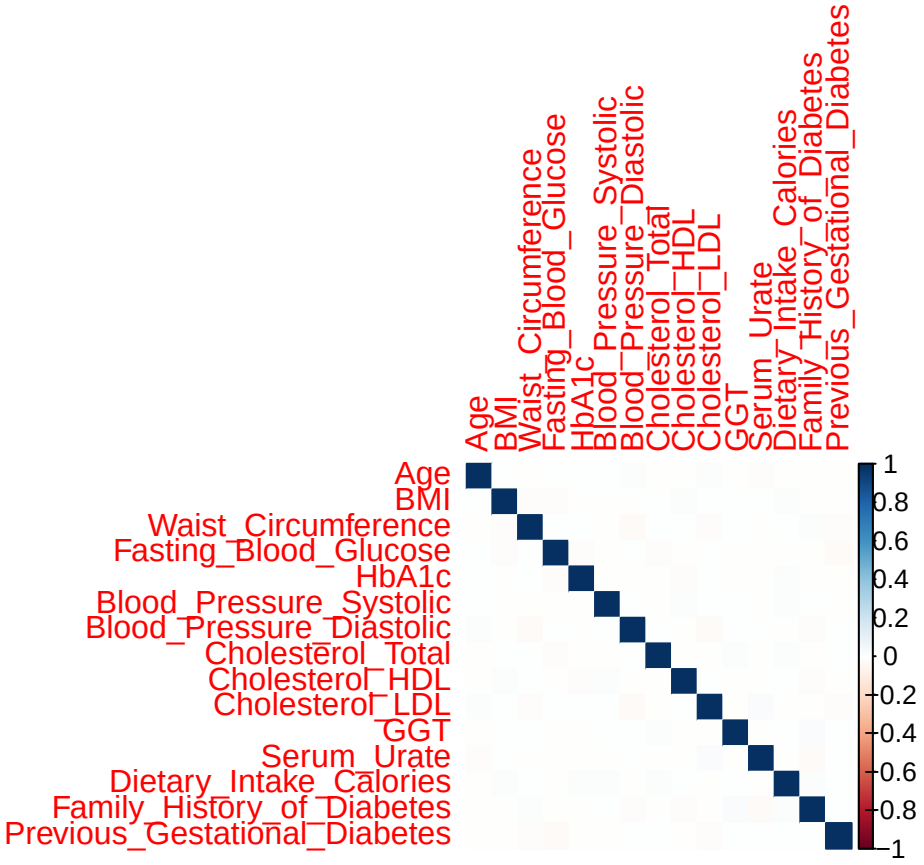
```
library(corrplot)
cor(diabetes_data2)
```

```
##                                     Age           BMI Waist_Circumference
## Age                          1.000000000  0.0093518945       -2.601642e-03
## BMI                          0.009351894  1.0000000000       -1.429071e-02
## Waist_Circumference         -0.002601642 -0.0142907106        1.000000e+00
## Fasting_Blood_Glucose        0.002696007 -0.0144780864        9.080049e-03
## HbA1c                        0.003153032 -0.0081629581        1.884745e-05
## Blood_Pressure_Systolic      0.002470229 -0.0022347931       -3.090026e-03
## Blood_Pressure_Diastolic     0.011472054 -0.0071627484       -2.035503e-02
## Cholesterol_Total           -0.005088170  0.0046856415        5.029529e-03
## Cholesterol_HDL             -0.005920353  0.0115938385        1.525451e-03
## Cholesterol_LDL              0.013345416  0.0003128711       -1.551238e-02
## GGT                         -0.001155861  0.0011051281        9.368938e-03
## Serum_Urate                 -0.014949828  0.0023757062       -3.502474e-03
## Dietary_Intake_Calories     -0.006664575  0.0178760804        2.031754e-03
## Family_History_of_Diabetes  -0.004412801 -0.0077284387        1.949711e-02
## Previous_Gestational_Diabetes -0.000353904 -0.0011300029       -1.516781e-02
##                             Fasting_Blood_Glucose         HbA1c
## Age                            0.0026960073  3.153032e-03
## BMI                           -0.0144780864 -8.162958e-03
## Waist_Circumference            0.0090800492  1.884745e-05
## Fasting_Blood_Glucose          1.0000000000 -1.355401e-02
## HbA1c                         -0.0135540144  1.000000e+00
## Blood_Pressure_Systolic       -0.0001392489 -4.640586e-03
## Blood_Pressure_Diastolic       0.0045557416 -8.879355e-03
## Cholesterol_Total             -0.0156488021 -4.754061e-03
## Cholesterol_HDL               -0.0021107816 -1.280672e-02
## Cholesterol_LDL                0.0075275868  7.864277e-03
## GGT                           -0.0052003569 -2.462741e-03
## Serum_Urate                   -0.0063361129 -1.841983e-03
## Dietary_Intake_Calories       -0.0067419487  1.189038e-02
## Family_History_of_Diabetes    -0.0006547506 -6.156078e-04
## Previous_Gestational_Diabetes -0.0234213672 -1.706655e-03
##                             Blood_Pressure_Systolic Blood_Pressure_Diastolic
## Age                              0.0024702291             0.0114720537
## BMI                             -0.0022347931            -0.0071627484
## Waist_Circumference             -0.0030900263            -0.0203550286
## Fasting_Blood_Glucose           -0.0001392489             0.0045557416
## HbA1c                           -0.0046405857            -0.0088793554
## Blood_Pressure_Systolic          1.0000000000             0.0009733392
## Blood_Pressure_Diastolic         0.0009733392             1.0000000000
## Cholesterol_Total               -0.0033663101             0.0039070103
## Cholesterol_HDL                  0.0106890228            -0.0048671800
## Cholesterol_LDL                  0.0071124398            -0.0211462980
```

```
## GGT                                    0.0012551640             0.0018586968
## Serum_Urate                           -0.0092861242             0.0043019707
## Dietary_Intake_Calories                0.0190296377            -0.0024800386
## Family_History_of_Diabetes             0.0021377167            -0.0121798009
## Previous_Gestational_Diabetes         -0.0089865481             0.0085713273
##                               Cholesterol_Total Cholesterol_HDL Cholesterol_LDL
## Age                              -0.005088170   -0.0059203530    0.0133454156
## BMI                               0.004685641    0.0115938385    0.0003128711
## Waist_Circumference               0.005029529    0.0015254512   -0.0155123767
## Fasting_Blood_Glucose            -0.015648802   -0.0021107816    0.0075275868
## HbA1c                            -0.004754061   -0.0128067197    0.0078642769
## Blood_Pressure_Systolic          -0.003366310    0.0106890228    0.0071124398
## Blood_Pressure_Diastolic          0.003907010   -0.0048671800   -0.0211462980
## Cholesterol_Total                 1.000000000   -0.0096787957   -0.0022409704
## Cholesterol_HDL                  -0.009678796    1.0000000000    0.0058508216
## Cholesterol_LDL                  -0.002240970    0.0058508216    1.0000000000
## GGT                               0.013313522   -0.0092052895   -0.0059889453
## Serum_Urate                      -0.004182229   -0.0080397915    0.0238858615
## Dietary_Intake_Calories           0.010119044    0.0007799387   -0.0074252200
## Family_History_of_Diabetes       -0.008553071   -0.0106631181   -0.0042041629
## Previous_Gestational_Diabetes     0.003354057   -0.0014026157   -0.0110075646
##                                          GGT   Serum_Urate
## Age                              -0.001155861 -0.0149498283
## BMI                               0.001105128  0.0023757062
## Waist_Circumference               0.009368938 -0.0035024738
## Fasting_Blood_Glucose            -0.005200357 -0.0063361129
## HbA1c                            -0.002462741 -0.0018419826
## Blood_Pressure_Systolic           0.001255164 -0.0092861242
## Blood_Pressure_Diastolic          0.001858697  0.0043019707
## Cholesterol_Total                 0.013313522 -0.0041822288
## Cholesterol_HDL                  -0.009205290 -0.0080397915
## Cholesterol_LDL                  -0.005988945  0.0238858615
## GGT                               1.000000000 -0.0021858839
## Serum_Urate                      -0.002185884  1.0000000000
## Dietary_Intake_Calories           0.003922087 -0.0003266329
## Family_History_of_Diabetes        0.025674539 -0.0227222245
## Previous_Gestational_Diabetes     0.005011470  0.0054577549
##                               Dietary_Intake_Calories
## Age                                     -0.0066645755
## BMI                                      0.0178760804
## Waist_Circumference                      0.0020317544
## Fasting_Blood_Glucose                   -0.0067419487
## HbA1c                                    0.0118903831
## Blood_Pressure_Systolic                  0.0190296377
## Blood_Pressure_Diastolic                -0.0024800386
## Cholesterol_Total                        0.0101190444
## Cholesterol_HDL                          0.0007799387
## Cholesterol_LDL                         -0.0074252200
## GGT                                      0.0039220874
## Serum_Urate                             -0.0003266329
## Dietary_Intake_Calories                  1.0000000000
## Family_History_of_Diabetes               0.0163686471
## Previous_Gestational_Diabetes            0.0090545061
##                               Family_History_of_Diabetes
```

```
## Age                                        -0.0044128013
## BMI                                        -0.0077284387
## Waist_Circumference                         0.0194971054
## Fasting_Blood_Glucose                      -0.0006547506
## HbA1c                                      -0.0006156078
## Blood_Pressure_Systolic                     0.0021377167
## Blood_Pressure_Diastolic                   -0.0121798009
## Cholesterol_Total                          -0.0085530708
## Cholesterol_HDL                            -0.0106631181
## Cholesterol_LDL                            -0.0042041629
## GGT                                         0.0256745388
## Serum_Urate                                -0.0227222245
## Dietary_Intake_Calories                     0.0163686471
## Family_History_of_Diabetes                  1.0000000000
## Previous_Gestational_Diabetes               0.0041406608
##                               Previous_Gestational_Diabetes
## Age                                          -0.000353904
## BMI                                          -0.001130003
## Waist_Circumference                          -0.015167812
## Fasting_Blood_Glucose                        -0.023421367
## HbA1c                                        -0.001706655
## Blood_Pressure_Systolic                      -0.008986548
## Blood_Pressure_Diastolic                      0.008571327
## Cholesterol_Total                             0.003354057
## Cholesterol_HDL                              -0.001402616
## Cholesterol_LDL                              -0.011007565
## GGT                                           0.005011470
## Serum_Urate                                   0.005457755
## Dietary_Intake_Calories                       0.009054506
## Family_History_of_Diabetes                    0.004140661
## Previous_Gestational_Diabetes                 1.000000000
```

```
corrplot(cor(diabetes_data2), method='color')
```



It appears that there are no strongly correlated variables and the PCA did not provide any conclusive results to help us pick which variables to use. For this reason, we will pick our own variables which will be: BMI, Age, Waist_Circumference, Cholesterol_LDL, and Family_History_of_Diabetes which gave us the highest accuracy after testing multiple combinations.

Here, we wanted to see how many observations there were of each HbA1c level to get an idea of the ranges recorded.

```
diabetes_Summary <- summarize(group_by(diabetes_data, HbA1c), Count=n(), Percentage=n()/nrow(diabetes_da
kable(diabetes_Summary)
```

HbA1c	Count	Percentage
4.0	59	0.59
4.1	100	1.00
4.2	88	0.88
4.3	89	0.89
4.4	95	0.95
4.5	94	0.94
4.6	82	0.82
4.7	85	0.85
4.8	81	0.81
4.9	94	0.94
5.0	93	0.93
5.1	88	0.88

HbA1c	Count	Percentage
5.2	97	0.97
5.3	80	0.80
5.4	88	0.88
5.5	82	0.82
5.6	76	0.76
5.7	103	1.03
5.8	87	0.87
5.9	95	0.95
6.0	101	1.01
6.1	95	0.95
6.2	80	0.80
6.3	80	0.80
6.4	104	1.04
6.5	82	0.82
6.6	93	0.93
6.7	98	0.98
6.8	92	0.92
6.9	87	0.87
7.0	97	0.97
7.1	105	1.05
7.2	109	1.09
7.3	84	0.84
7.4	88	0.88
7.5	86	0.86
7.6	96	0.96
7.7	86	0.86
7.8	101	1.01
7.9	102	1.02
8.0	91	0.91
8.1	75	0.75
8.2	79	0.79
8.3	95	0.95
8.4	90	0.90
8.5	91	0.91
8.6	93	0.93
8.7	77	0.77
8.8	99	0.99
8.9	72	0.72
9.0	76	0.76
9.1	94	0.94
9.2	83	0.83
9.3	88	0.88
9.4	97	0.97
9.5	89	0.89
9.6	82	0.82
9.7	96	0.96
9.8	88	0.88
9.9	95	0.95
10.0	78	0.78
10.1	89	0.89
10.2	99	0.99
10.3	86	0.86

HbA1c	Count	Percentage
10.4	88	0.88
10.5	77	0.77
10.6	91	0.91
10.7	96	0.96
10.8	87	0.87
10.9	119	1.19
11.0	85	0.85
11.1	103	1.03
11.2	89	0.89
11.3	100	1.00
11.4	88	0.88
11.5	79	0.79
11.6	95	0.95
11.7	115	1.15
11.8	90	0.90
11.9	96	0.96
12.0	86	0.86
12.1	94	0.94
12.2	83	0.83
12.3	90	0.90
12.4	103	1.03
12.5	102	1.02
12.6	78	0.78
12.7	112	1.12
12.8	83	0.83
12.9	103	1.03
13.0	90	0.90
13.1	81	0.81
13.2	74	0.74
13.3	101	1.01
13.4	81	0.81
13.5	113	1.13
13.6	93	0.93
13.7	92	0.92
13.8	100	1.00
13.9	91	0.91
14.0	97	0.97
14.1	81	0.81
14.2	84	0.84
14.3	96	0.96
14.4	87	0.87
14.5	85	0.85
14.6	95	0.95
14.7	78	0.78
14.8	98	0.98
14.9	91	0.91
15.0	36	0.36

K-Nearest Neighbors Clustering

In order to successfully do k-Nearest Neighbors Clustering, we have to create a categorical variable from the data set. For all observations between 0 and 5.7 they are classified as normal, all observations between 5.7 and 6.4 are pre-diabetic, and all observations over 6.4 are classified as diabetic.

```
diabetes_data2$HbA1c_levels <- cut(diabetes_data2$HbA1c,
                      breaks = c(0, 5.7, 6.4, Inf),
                      labels = c("Normal", "Pre-diabetic", "Diabetic"))
diabetes_data2$HbA1c_levels <- as.factor(diabetes_data2$HbA1c_levels)
```

Here is a table of the number of observations in each category.

```
table(diabetes_data2$HbA1c_levels)
```

```
##
##       Normal Pre-diabetic     Diabetic
##         1574          642         7784
```

Based off of this categorization, we can compare this directly to the confusion matrix after we use kNN clustering to test accuracy.

```
X_diabetes <- diabetes_data[ , c("BMI", "Age", "Waist_Circumference", "Cholesterol_LDL", "Family_History
y_diabetes <- diabetes_data2$HbA1c_levels
```

Here we are splitting the data into "test" and "training" data.

```
set.seed(234)
inTrain <- createDataPartition(y = y_diabetes, p = 0.75, list = FALSE)
```

```
X_train <- scale(X_diabetes[inTrain, ])
y_train <- y_diabetes[inTrain]

X_test <- scale(X_diabetes[-inTrain, ]) #the minus excludes the values specified
y_test <- y_diabetes[-inTrain]
```

Here I will test for the best number k to use.

```
set.seed(456)
diabetes_pred <- knn(train = X_train, test = X_test, cl = y_train, k=1)
fit <- train(x = X_train,
             y = y_train,
             method = "knn",
             tuneGrid = data.frame(k=c(1:25)),
             trControl = trainControl(method = "cv", number = 5),
             metric = "Accuracy")
fit
```

```
## k-Nearest Neighbors
##
## 7501 samples
```

```
##    5 predictor
##    3 classes: 'Normal', 'Pre-diabetic', 'Diabetic'
##
## No pre-processing
## Resampling: Cross-Validated (5 fold)
## Summary of sample sizes: 6002, 6001, 6001, 6000, 6000
## Resampling results across tuning parameters:
##
##   k   Accuracy   Kappa
##    1  0.6301822  -6.254721e-05
##    2  0.6296474   5.580711e-04
##    3  0.7053713  -7.484355e-04
##    4  0.7160398  -1.193879e-02
##    5  0.7400355  -6.146240e-03
##    6  0.7508338  -5.052992e-03
##    7  0.7608321  -4.897174e-03
##    8  0.7622980  -1.198718e-02
##    9  0.7677648  -7.333153e-03
##   10  0.7686973  -8.291618e-03
##   11  0.7732304  -4.350672e-03
##   12  0.7732301  -5.761494e-03
##   13  0.7749638  -4.409247e-03
##   14  0.7756299  -3.351595e-03
##   15  0.7765631  -2.515424e-03
##   16  0.7765631  -2.575791e-03
##   17  0.7769631  -2.266158e-03
##   18  0.7772296  -1.444019e-03
##   19  0.7777631  -1.618208e-04
##   20  0.7781630   1.444066e-04
##   21  0.7780296  -8.301144e-05
##   22  0.7780297  -4.546563e-04
##   23  0.7782963   7.474784e-04
##   24  0.7782964   7.476300e-04
##   25  0.7784296   9.748823e-04
##
## Accuracy was used to select the optimal model using the largest value.
## The final value used for the model was k = 25.
```

```
diabetes_pred <- knn(train = X_train, test = X_test, cl = y_train, k=25)
CrossTable(x = y_test, y = diabetes_pred, prop.chisq = FALSE)
```

```
##
##
##    Cell Contents
## |-------------------------|
## |                       N |
## |         N / Table Total |
## |-------------------------|
##
##
## Total Observations in Table:  2499
##
##
##              | diabetes_pred
```

```
##       y_test |  Diabetic | Row Total |
## -------------|-----------|-----------|
##       Normal |       393 |       393 |
##              |     0.157 |           |
## -------------|-----------|-----------|
## Pre-diabetic |       160 |       160 |
##              |     0.064 |           |
## -------------|-----------|-----------|
##     Diabetic |      1946 |      1946 |
##              |     0.779 |           |
## -------------|-----------|-----------|
## Column Total |      2499 |      2499 |
## -------------|-----------|-----------|
##
##
```

```
mean(diabetes_pred==y_test)
```

```
## [1] 0.7787115
```

```
mean(y_diabetes==0)
```

```
## [1] 0
```

Despite the fairly high accuracy, the dataset is unbalanced with significantly more diabetics than not which is causing the program to classify all the observations as diabetic. So, we will also do k-means clustering within the diabetics alone to try and see if we can differentiate between type 1 and type 2 diabetics. To do so, I will make a new variable with only the diabetics.

```
diabetes_data$HbA1c_levels <- cut(diabetes_data2$HbA1c,
                      breaks = c(6.4, Inf),
                      labels = c("Diabetic"))
diabetes_data$HbA1c_levels <- as.factor(diabetes_data$HbA1c_levels)
```

```
diabetes_data2$HbA1c <- as.numeric(as.character(diabetes_data2$HbA1c))
diabetes_data2_filtered <- subset(diabetes_data2, HbA1c >= 6.4)
```

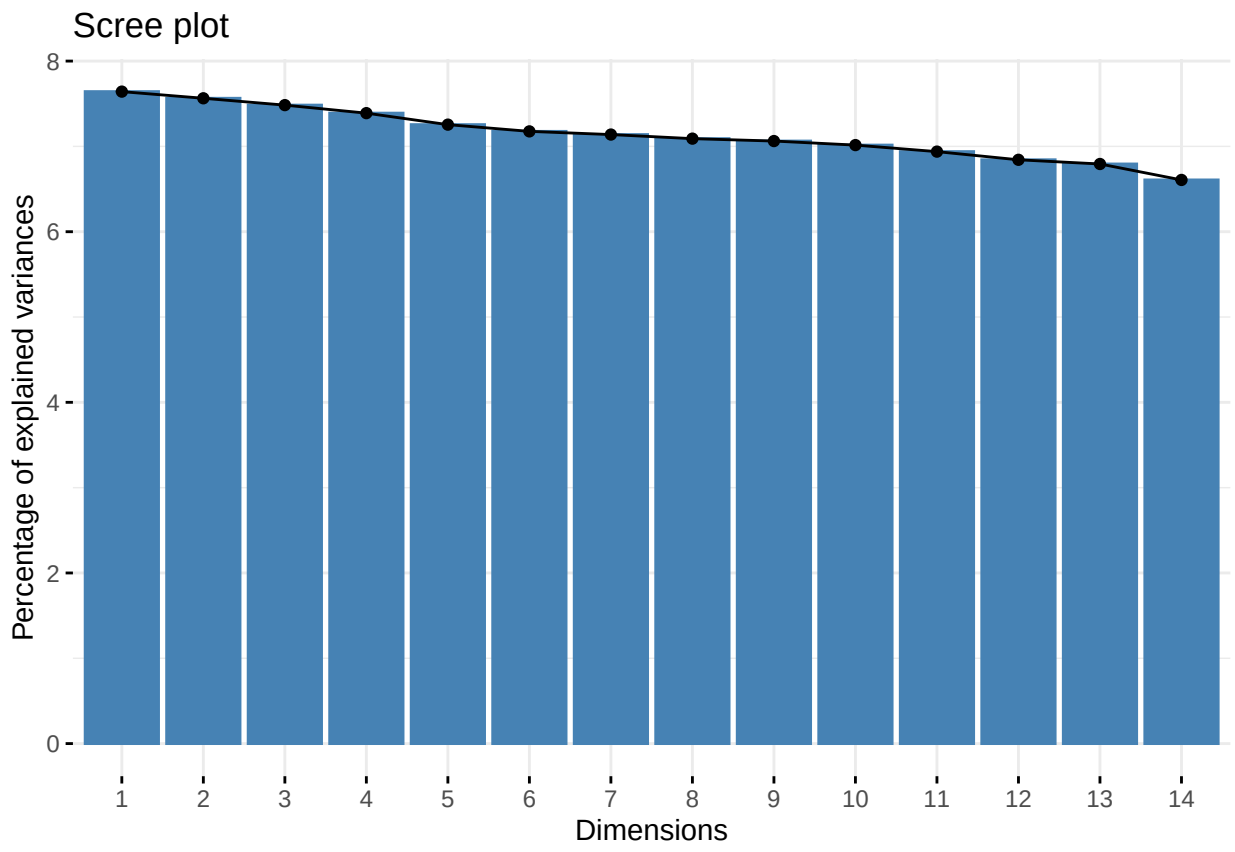
```
diabetic_data <- subset(diabetes_data2_filtered, HbA1c_levels == "Diabetic")
diabetic_numeric <- diabetic_data %>%
  select(-HbA1c_levels, -HbA1c) %>%
  select(where(is.numeric))

pr.out <- prcomp(diabetic_numeric, scale = TRUE)
summary(pr.out)
```

```
## Importance of components:
##                           PC1     PC2     PC3    PC4     PC5     PC6     PC7
## Standard deviation     1.03435 1.02906 1.02359 1.0171 1.00782 1.00232 0.99971
## Proportion of Variance 0.07642 0.07564 0.07484 0.0739 0.07255 0.07176 0.07139
## Cumulative Proportion  0.07642 0.15206 0.22690 0.3008 0.37334 0.44510 0.51649
##                           PC8     PC9    PC10    PC11    PC12    PC13    PC14
```

```
## Standard deviation     0.99636 0.99436 0.99102 0.98560 0.97878 0.97526 0.96175
## Proportion of Variance 0.07091 0.07063 0.07015 0.06939 0.06843 0.06794 0.06607
## Cumulative Proportion  0.58740 0.65803 0.72818 0.79756 0.86599 0.93393 1.00000
```

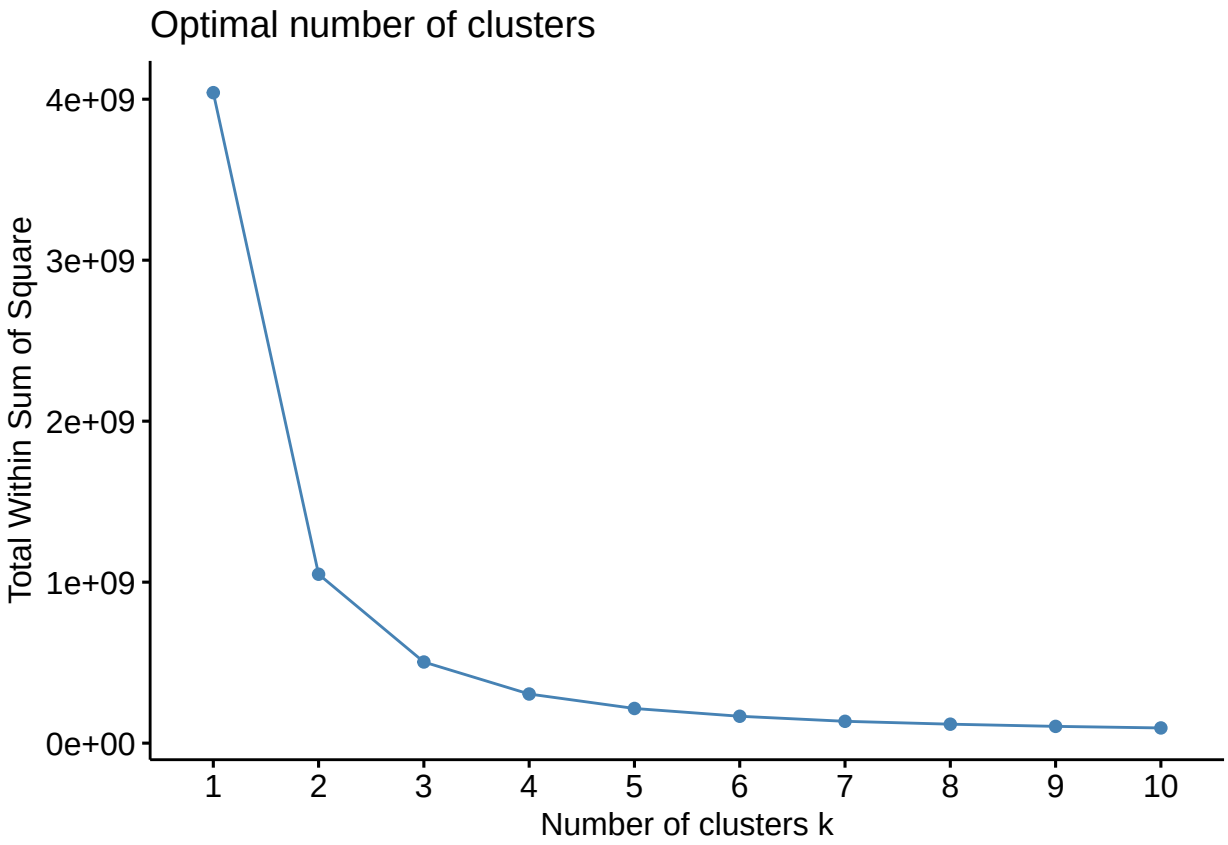
```
fviz_screeplot(pr.out, ncp = ncol(diabetic_numeric))
```



K-Means Clustering

As you can see, the screeplot is also inconclusive as all variables are very close together. For this reason, we did outside research to determine the variable we will use for k-means (Cholesterol_HDL). We will make a scree-plot to use the elbow method to determine which variables to use for k-means clustering using only the diabetics.

```
set.seed(789)
fviz_nbclust(diabetic_numeric, kmeans, method = "wss", k.max = 10)
```



After doing some research, I discovered that HDL cholesterol (the "good cholesterol") can be an indicator of Type 1 diabetes because insulin plays a crucial role in your HDL cholesterol metabolism. In individuals who are insulin deficient, such as type 1 diabetics, HDL tends to be higher to compensate for the lack of insulin. For this reason, I tested k-means clustering against HbA1c levels (key indicator for all diabetics) and HDL cholesterol levels. We scaled the data to make it usable for k-means clustering.

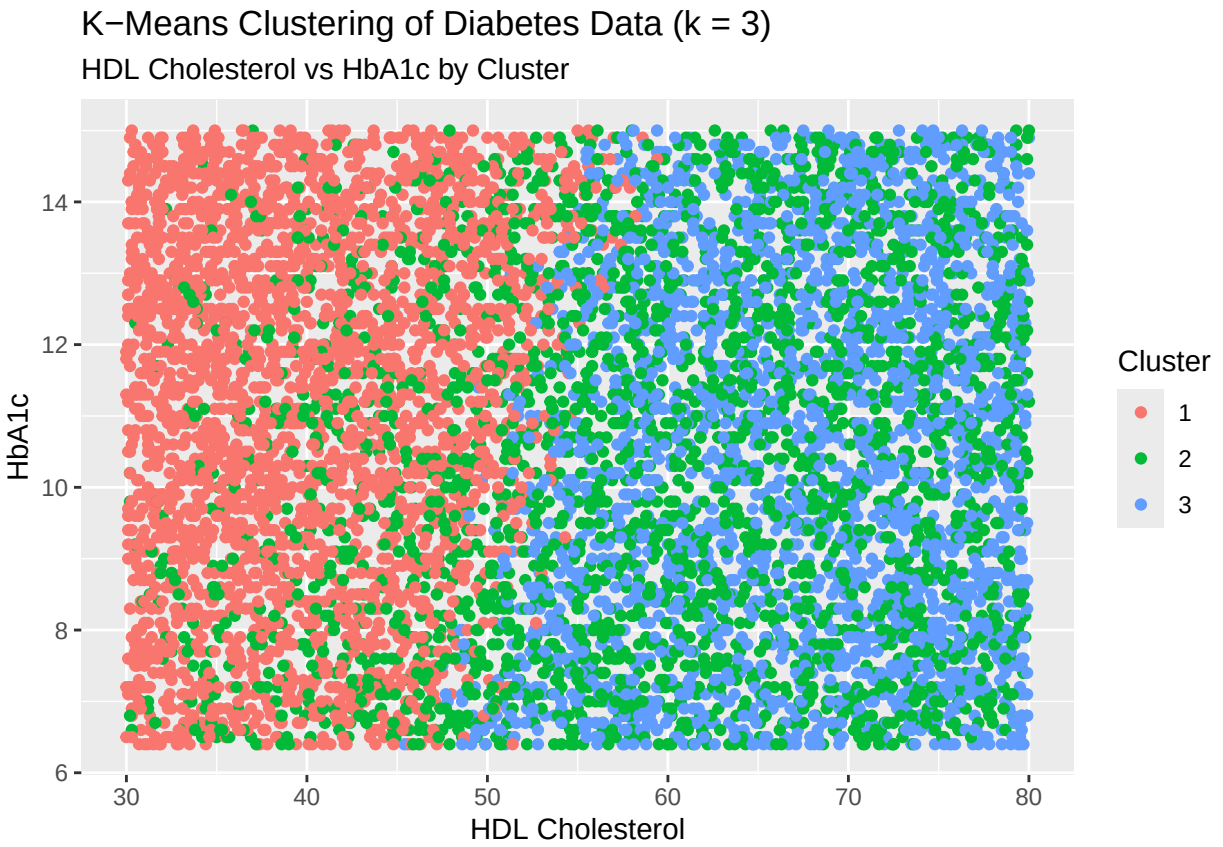
```
diabetic_numeric <- select(diabetes_data2_filtered, BMI, Age, Waist_Circumference, Cholesterol_LDL, Cho

scaled_diabetic_numeric <- scale(select(diabetic_numeric, -Family_History_of_Diabetes))

set.seed(123)
kmeans_result <- kmeans(scaled_diabetic_numeric, centers = 3)

diabetic_numeric$Cluster <- as.factor(kmeans_result$cluster)

ggplot(diabetic_numeric, aes(x = Cholesterol_HDL, y = HbA1c, color = Cluster)) +
  geom_point() +
  labs(title = "K-Means Clustering of Diabetes Data (k = 3)",
       subtitle = "HDL Cholesterol vs HbA1c by Cluster",
       x = "HDL Cholesterol",
       y = "HbA1c")
```



Conclusion

Since type 1 diabetics tend to have higher levels of HDL cholesterol, the blue cluster is likely the type 1 diabetics. The type 2 diabetics tend to have lower levels of HDL cholesterol because they are not insulin deficient, so the body does not over compensate with HDL like insulin deficient people do. Regardless, HDL cholesterol tends to differ in both types of diabetics so there is an indeterminate cluster (the green one) that is mixed with both types.

Overall, this data set was very interesting to work with, but left us with inconclusive results. None of the measurements, aside from HbA1c levels, allowed for any kind of accurate classification. While there were somewhat distinct clusters of type 1 and type 2 based on HDL levels, there is no way of knowing from this data if it is correct since the HDL cholesterol is not a determining factor of the type. Instead, the type is usually determined from tests on autoantibody tests and C-peptide levels, and neither of which were measured in this data set. Regardless, using what we had, I enjoyed testing different methods we used in class for different things using this data. For the future, I would have liked to have a more definitive way of classifying things, and perhaps different methods to test for relationships between different factors. I do wonder why certain things were recorded (such as smoking status) since these typically do not have much relevance to diabetes diagnoses.