DSCI 5300 Final Project

Liam Frank

2023-10-20

Introduction

With the recent rise in healthcare analytics and data-driven decision making, this study aims to answer the question, "can the crude prevalence of cancer be accurately predicted and classified on a city by city basis?" In an attempt to answer the proposed question, this study will be conducted using a variety of linear regression models, a support vector machine classification, and k means clustering. Crude prevalence is a ratio of number of cases or deaths in a specified population per year, usually expressed as the number of cases per 100,000. The potential benefits of being able to accurately predict the prude prevalence of cancer include optimal resource allocation, healthcare planning, potential cost savings to both patients and hospitals, early detection, and an overall improved quality of life. Using the power of data, this study looks to answer the proposed question and contribute to the listed benefits of accurate crude prevalence predictions.

The Data

The Data Set is comprised of a total of 198 different cities and 28 variables. The populations of the cities included range from 122 to 11,408, with a median of value of 3,812 and a mean value of 3,992. Other than the population variable, the 27 remaining variables are all in respect to measurements of crude prevalence for various different diseases, and common health issues.

```r
health<-read.csv("Health of CityX(1).csv")
str(health)
```

```
## 'data.frame':    198 obs. of  28 variables:
##  $ Population2010        : int  3611 2552 1546 3009 3394 2687 4939 3045 3223 2697 ...
##  $ ACCESS2_CrudePrev     : num  10 8.2 15.5 12.3 13.4 11.1 13.6 14.4 13.8 16.5 ...
##  $ ARTHRITIS_CrudePrev   : num  19 23.4 8.3 13.5 12 18.5 10.9 9.6 10.6 11.7 ...
##  $ BINGE_CrudePrev       : num  24 20.8 27 27.1 28.9 24.4 28.7 29.6 29.7 26.6 ...
##  $ BPHIGH_CrudePrev      : num  23.8 28.7 13.4 18.3 17.2 23.9 15.7 14.8 15.7 18 ...
##  $ BPMED_CrudePrev       : num  73.9 77.9 50.7 63.6 60.7 72.3 58.7 55 56.3 59.9 ...
##  $ CANCER_CrudePrev      : num  7.5 8.8 2.2 4.1 3.8 6.4 3.3 2.9 3.2 3.3 ...
##  $ CASTHMA_CrudePrev     : num  7.3 7.3 8.8 7.9 7.5 7.6 7.7 7.6 7.9 7.7 ...
##  $ CHD_CrudePrev         : num  4.5 5.4 1.9 2.9 2.5 4.1 2.4 2.1 2.3 2.4 ...
##  $ CHECKUP_CrudePrev     : num  66.1 70.1 57.5 60.5 60 65.1 58.8 57.9 57.8 59.8 ...
##  $ CHOLSCREEN_CrudePrev  : num  83.7 88.8 65.9 76.3 77.8 83.3 72.9 72.2 72.5 76.6 ...
##  $ COLON_SCREEN_CrudePrev: num  72 74.6 64.1 66.8 66.8 70.3 67.2 65.3 64.4 65.6 ...
##  $ COPD_CrudePrev        : num  3.4 4.1 2.6 3.1 2.6 3.6 2.7 2.4 2.7 2.7 ...
##  $ COREM_CrudePrev       : num  42.4 43.8 35.5 38.3 39.1 40.5 37.2 37.4 36.5 35 ...
##  $ COREW_CrudePrev       : num  30.7 34.9 27.7 30.1 30.7 30.6 29.5 27.4 28.7 28.5 ...
##  $ CSMOKING_CrudePrev    : num  8.6 8.2 11.8 11.3 10.9 10.3 11.3 10.9 12.2 12.3 ...
##  $ DENTAL_CrudePrev      : num  79.2 81.9 69.3 74.7 74 76.8 72.9 72.5 71.1 70 ...
##  $ DIABETES_CrudePrev    : num  6.2 8 3.6 5.2 4.8 6.9 4.3 4.1 4.2 5.4 ...
##  $ HIGHCHOL_CrudePrev    : num  30.2 36.2 17.8 25.1 23.7 30.5 21.9 20.4 21.6 23.5 ...
##  $ KIDNEY_CrudePrev      : num  2.3 2.5 1.4 1.8 1.6 2.2 1.5 1.4 1.5 1.7 ...
##  $ LPA_CrudePrev         : num  15.8 15.9 17.5 16.8 15.9 17 16.3 16 16.6 18.5 ...
##  $ MAMMOUSE_CrudePrev    : num  80.5 80.4 80.8 80.4 80.7 80.2 80.4 81.2 80.2 81.7 ...
##  $ MHLTH_CrudePrev       : num  8.3 7.8 13.7 11 9.9 9.3 11 10.9 11.6 10.9 ...
##  $ OBESITY_CrudePrev     : num  19.9 21.4 20.2 21.8 21.3 22.4 20.2 20.6 21.5 24 ...
##  $ PHLTH_CrudePrev       : num  7.1 7.9 6.4 7.1 6.4 7.6 6.3 6 6.6 7.1 ...
##  $ SLEEP_CrudePrev       : num  24.4 23.6 26.7 26.4 27 26 26.7 26.3 27.3 28.8 ...
##  $ STROKE_CrudePrev      : num  2.1 2.4 1 1.5 1.3 2 1.2 1.1 1.2 1.4 ...
##  $ TEETHLOST_CrudePrev   : num  5.7 4.4 8.9 6.6 6.3 6 6.6 6.5 7.6 7.6 ...
```

Libraries Used

```r
library(corrplot)
library(ggplot2)
library(e1071)
library(Metrics)
library(MASS)
library(cluster)
library(factoextra)
```

GGplot2 and corrplot were both used for exploratory analysis. Metrics was used to calculate mean squared error and mean absolute error for the regression models. MASS was used for the boxcox transformation analysis. e1071 was used for creating the support vector machine classification model. Cluster was used for the k means clustering. Factoextra was used for creating visualizations of k means clustering.

Data Cleaning

```r
sum(is.na(health))
```

```
## [1] 2
```

```r
healthnew<-na.omit(health)
```

The only cleaning that originally took place was omitting the two observations that contained NA values. This decreases the data set to 196 total observations. Later transformation and cleaning was done in regards converting CANCER_CrudePrev to a factor variable that allowed for classification using a support vector machine. That transformation will be discussed in the respective section of the paper.

Train and Test Set Creation

```r
set.seed(123)
extrain<- sample(nrow(healthnew), 0.70*nrow(healthnew), replace=FALSE)
healthTrain <- healthnew[extrain,]
healthTest <-  healthnew[-extrain,]
```

A standard 70:30 split was used to divide the data into a train and test set. This was done in order to test the created models and minimize the possibility of over fitting. Set.seed was used for re-producibility of results.

Exploratory Analysis

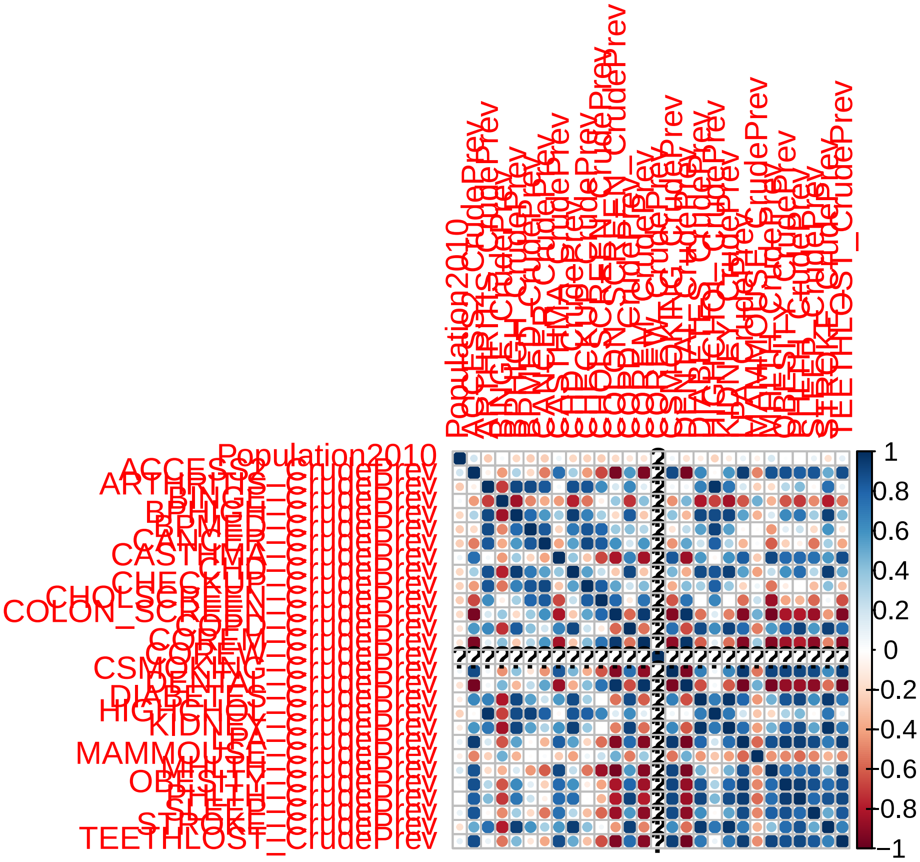
```
summary(healthnew$CANCER_CrudePrev)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   0.900   3.400   4.300   4.636   5.625  10.200
```

```
summary(healthnew)
```

```
##  Population2010  ACCESS2_CrudePrev ARTHRITIS_CrudePrev BINGE_CrudePrev
##  Min.   :  122   Min.   : 7.50     Min.   : 5.30       Min.   :15.60
##  1st Qu.: 2694   1st Qu.:12.47     1st Qu.:13.38       1st Qu.:21.70
##  Median : 3812   Median :17.85     Median :16.05       Median :23.70
##  Mean   : 3984   Mean   :20.61     Mean   :16.06       Mean   :23.78
##  3rd Qu.: 5178   3rd Qu.:27.80     3rd Qu.:18.70       3rd Qu.:25.80
##  Max.   :11408   Max.   :44.00     Max.   :33.10       Max.   :30.80
##  BPHIGH_CrudePrev BPMED_CrudePrev CANCER_CrudePrev CASTHMA_CrudePrev
##  Min.   :11.20    Min.   :24.50   Min.   : 0.900   Min.   : 6.100
##  1st Qu.:20.20    1st Qu.:61.80   1st Qu.: 3.400   1st Qu.: 7.500
##  Median :23.25    Median :67.05   Median : 4.300   Median : 7.900
##  Mean   :23.31    Mean   :65.95   Mean   : 4.636   Mean   : 8.074
##  3rd Qu.:26.43    3rd Qu.:71.53   3rd Qu.: 5.625   3rd Qu.: 8.500
##  Max.   :44.50    Max.   :80.70   Max.   :10.200   Max.   :10.900
##  CHD_CrudePrev    CHECKUP_CrudePrev CHOLSCREEN_CrudePrev COLON_SCREEN_CrudePrev
##  Min.   : 1.300   Min.   :54.10     Min.   :48.00        Min.   :38.20
##  1st Qu.: 2.975   1st Qu.:60.00     1st Qu.:73.78        1st Qu.:56.60
##  Median : 3.700   Median :62.65     Median :77.90        Median :64.90
##  Mean   : 3.913   Mean   :62.72     Mean   :77.48        Mean   :62.82
##  3rd Qu.: 4.800   3rd Qu.:65.10     3rd Qu.:81.65        3rd Qu.:69.45
##  Max.   :11.800   Max.   :72.40     Max.   :89.60        Max.   :75.50
##  COPD_CrudePrev   COREM_CrudePrev COREW_CrudePrev CSMOKING_CrudePrev
##  Min.   : 1.600   Min.   :20.10   Min.   :14.60   Min.   : 7.20
##  1st Qu.: 3.175   1st Qu.:28.88   1st Qu.:21.82   1st Qu.:10.97
##  Median : 3.800   Median :35.45   Median :28.55   Median :13.45
##  Mean   : 4.041   Mean   :34.32   Mean   :27.85   Mean   :14.19
##  3rd Qu.: 4.800   3rd Qu.:40.02   3rd Qu.:32.90   3rd Qu.:17.12
##  Max.   :14.300   Max.   :45.40   Max.   :42.30   Max.   :30.90
##  DENTAL_CrudePrev DIABETES_CrudePrev HIGHCHOL_CrudePrev KIDNEY_CrudePrev
##  Min.   :30.10    Min.   : 2.600     Min.   :11.90      Min.   :1.100
##  1st Qu.:55.70    1st Qu.: 6.200     1st Qu.:25.40      1st Qu.:1.800
##  Median :69.35    Median : 7.500     Median :28.65      Median :2.200
##  Mean   :65.84    Mean   : 7.956     Mean   :28.34      Mean   :2.279
##  3rd Qu.:76.40    3rd Qu.: 9.525     3rd Qu.:31.32      3rd Qu.:2.625
##  Max.   :83.80    Max.   :20.400     Max.   :44.60      Max.   :5.000
##  LPA_CrudePrev   MAMMOUSE_CrudePrev MHLTH_CrudePrev  OBESITY_CrudePrev
##  Min.   :13.40   Min.   :74.00      Min.   : 7.400   Min.   :17.40
##  1st Qu.:17.18   1st Qu.:78.60      1st Qu.: 9.775   1st Qu.:22.05  
##  Median :21.25   Median :79.65      Median :11.400   Median :24.70  
##  Mean   :22.60   Mean   :79.41      Mean   :12.056   Mean   :25.92  
##  3rd Qu.:27.30   3rd Qu.:80.40      3rd Qu.:14.025   3rd Qu.:29.40  
##  Max.   :42.50   Max.   :83.10      Max.   :22.900   Max.   :40.20  
##  PHLTH_CrudePrev  SLEEP_CrudePrev STROKE_CrudePrev TEETHLOST_CrudePrev
##  Min.   : 4.600   Min.   :22.60   Min.   :0.800    Min.   : 3.300     
##  1st Qu.: 7.200   1st Qu.:27.00   1st Qu.:1.600    1st Qu.: 5.975     
##  Median : 8.600   Median :29.50   Median :1.900    Median : 8.600     
##  Mean   : 9.398   Mean   :29.68   Mean   :2.086    Mean   :10.946     
##  3rd Qu.:11.100   3rd Qu.:32.40   3rd Qu.:2.425    3rd Qu.:14.400     
##  Max.   :24.200   Max.   :38.00   Max.   :6.400    Max.   :40.900
```

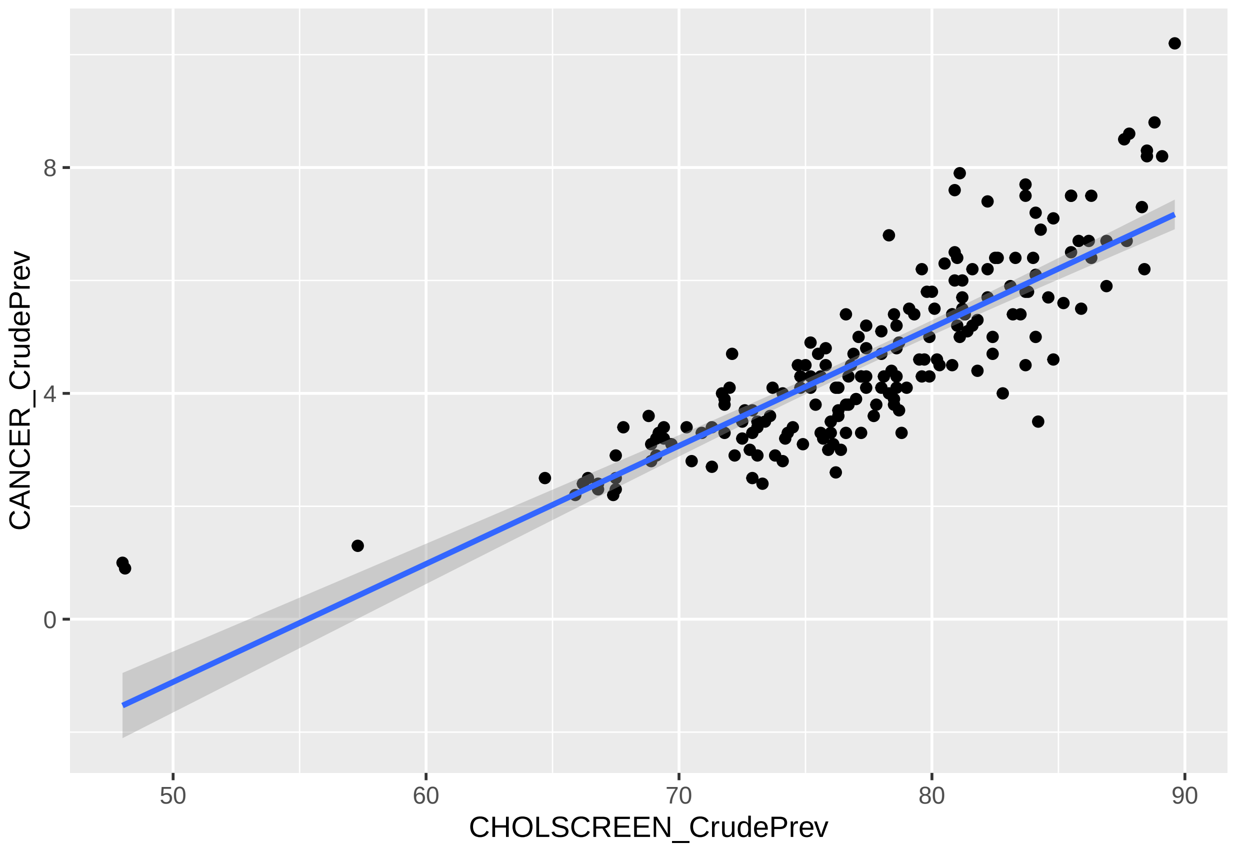
```
par(mfrow=c(1,1))
corrplot(cor(health))
```



```
cor(healthnew$CANCER_CrudePrev,healthnew$CHOLSCREEN_CrudePrev)
```

```
## [1] 0.8400612
```

```
ggplot(healthnew, aes(x=CHOLSCREEN_CrudePrev, y=CANCER_CrudePrev))+geom_point()+geom_smooth(method = 'l
```



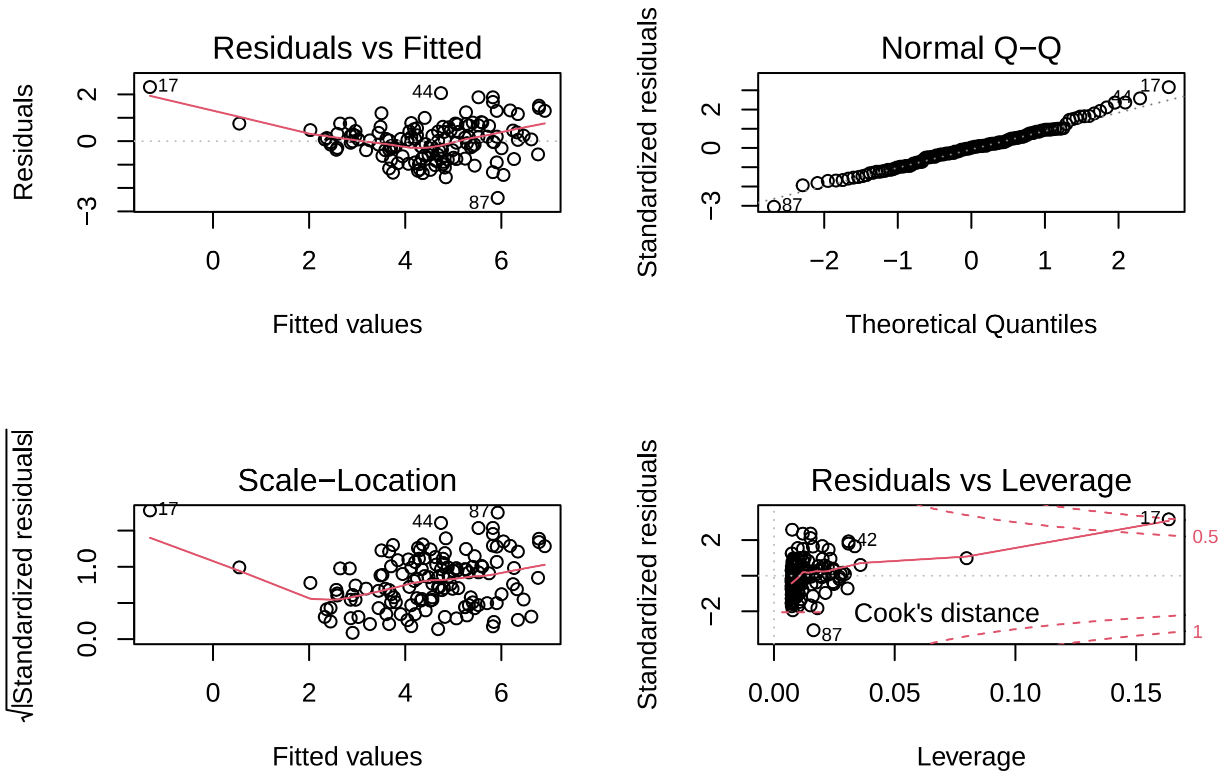
For exploratory analysis, a combination of correlations, correlation plots, descriptive statistics, and graphing of variable relationships were examined, not all of which have been included in the paper. From exploratory analysis it was found that the crude prevalence of cancer varies widely by city, suggesting that it would be a more than adequate prediction variable. For possible explanatory variables to predict the crude prevalence of cancer, correlation was used. As seen from the results, cholesterol crude prevalence had a high correlation factor to cancer crude prevalence.

Single Variable Linear Regression

```r
cancerLM1<-lm(CANCER_CrudePrev ~ CHOLSCREEN_CrudePrev, data =healthTrain)
summary(cancerLM1)
```

```
##
## Call:
## lm(formula = CANCER_CrudePrev ~ CHOLSCREEN_CrudePrev, data = healthTrain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.42440 -0.56397  0.01498  0.45511  2.31193
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)          -10.90707    0.83933  -12.99   <2e-16 ***
## CHOLSCREEN_CrudePrev   0.19990    0.01084   18.44   <2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.8 on 135 degrees of freedom
## Multiple R-squared:  0.7158, Adjusted R-squared:  0.7137
## F-statistic:   340 on 1 and 135 DF,  p-value: < 2.2e-16
```

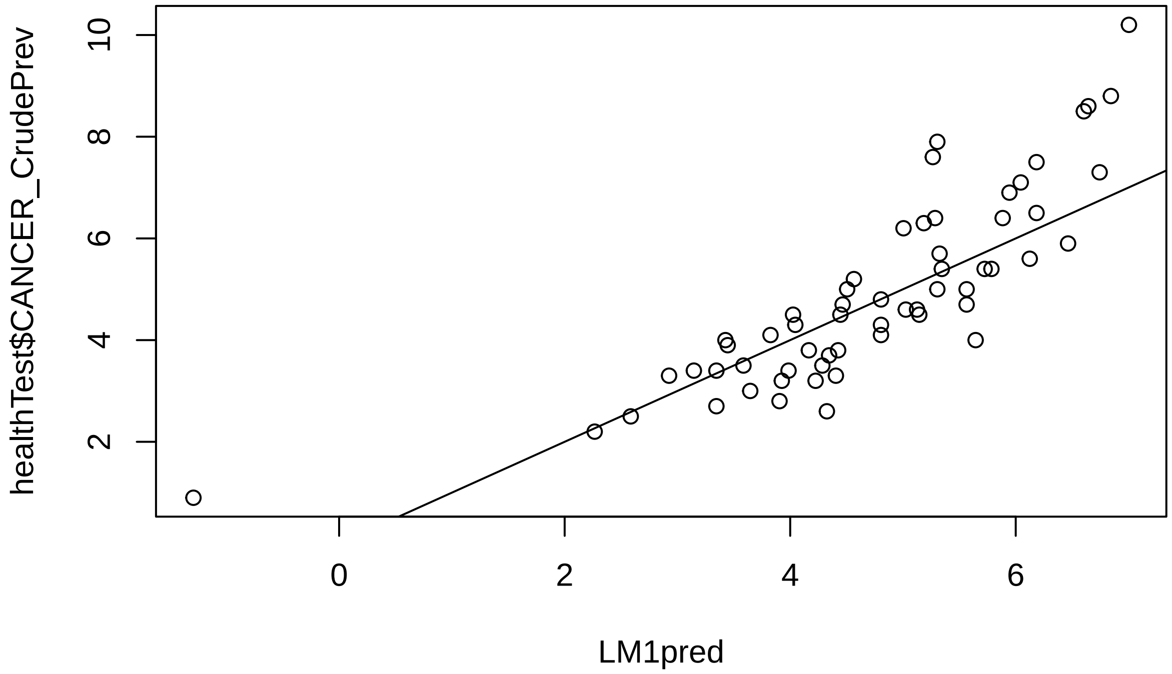
```
par(mfrow=c(2,2))
plot(cancerLM1)
```



```
LM1pred <- predict(cancerLM1, newdata = healthTest)
LM1data<- data.frame(actual=healthTest$CANCER_CrudePrev, predicted= LM1pred)
head(LM1data)
```

```
##    actual predicted
## 2     8.8  6.843930
## 3     2.2  2.266251
## 10    3.3  4.405167
## 11    5.0  5.304711
## 16    0.9 -1.291944
## 19    6.2  5.004863
```

```
par(mfrow=c(1,1))
plot(LM1pred,healthTest$CANCER_CrudePrev)+abline(a=0,b=1)
```



```
## integer(0)
```

```
mae(healthTest$CANCER_CrudePrev,LM1pred)
```

```
## [1] 0.7972689
```

```
mse(healthTest$CANCER_CrudePrev,LM1pred)
```

```
## [1] 1.099112
```

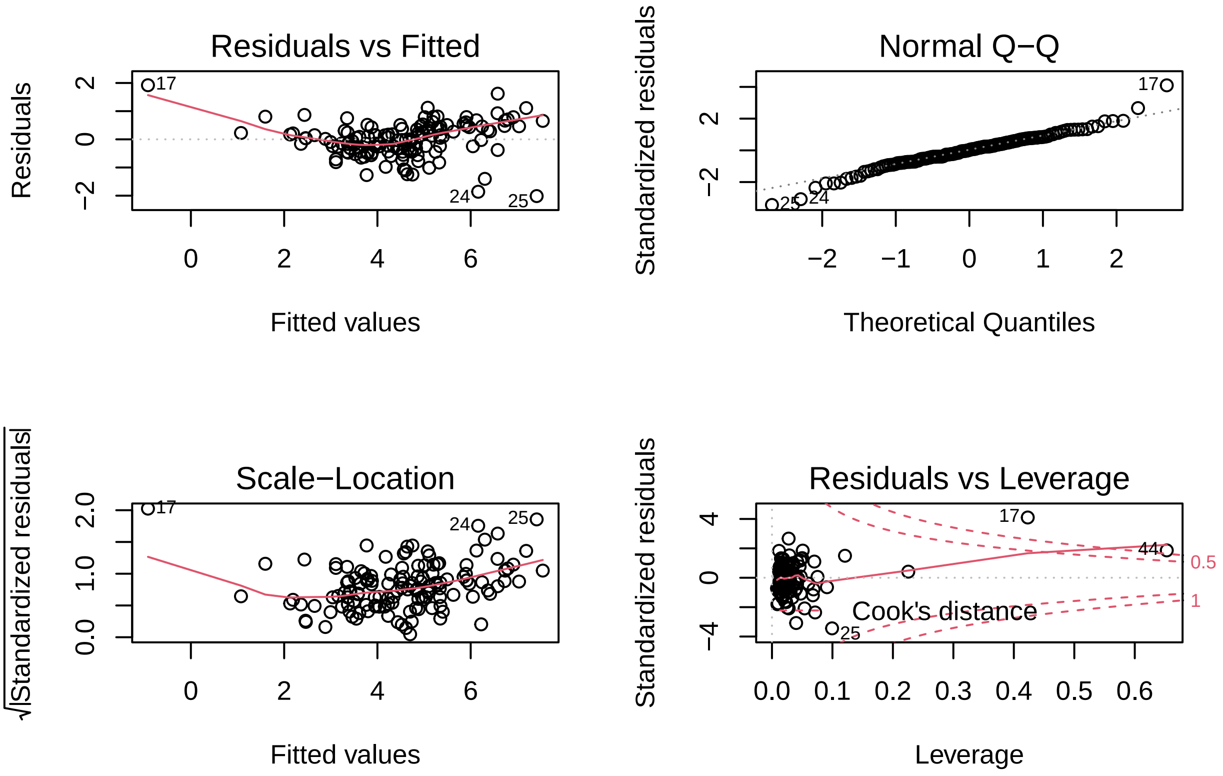
As seen in the summary, the model had a residual standard error of 0.8 with an adjusted R squared value of 0.71. Signifying that 71% of the variance in the crude prevalence of cancer can be attributed to crude prevalence of cholesterol screening. The coefficients and model are both significant, with P values less than 0.05. The residuals are evenly centered around zero, and the residuals vs fitted is evenly distributed as well as absent of a pattern.The mean squared error was 1.099 and the mean absolute error was 0.79. The normal qq plot follows the plotted line. In the predicted vs actual plot using the test set, the 45 degree abline follows linear nature of the predicted data points, but the predicted values contain a vast amount of variation.

Multiple Linear Regression Model

```
cancerLM2<-lm(CANCER_CrudePrev ~ COPD_CrudePrev+BPMED_CrudePrev+CHD_CrudePrev+CHOLSCREEN_CrudePrev+CHEC
summary(cancerLM2)
```

```
##
## Call:
## lm(formula = CANCER_CrudePrev ~ COPD_CrudePrev + BPMED_CrudePrev +
##     CHD_CrudePrev + CHOLSCREEN_CrudePrev + CHECKUP_CrudePrev -
##     1, data = healthTrain)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -2.01480 -0.34696  0.01575  0.39628  1.92044
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## COPD_CrudePrev       -1.30752    0.12577  -10.40  < 2e-16 ***
## BPMED_CrudePrev      -0.05641    0.02821   -2.00 0.047540 *
## CHD_CrudePrev         2.01898    0.18578   10.87  < 2e-16 ***
## CHOLSCREEN_CrudePrev  0.14748    0.02561    5.76 5.64e-08 ***
## CHECKUP_CrudePrev    -0.09108    0.02409   -3.78 0.000236 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.6167 on 132 degrees of freedom
## Multiple R-squared:  0.9838, Adjusted R-squared:  0.9832
## F-statistic:  1604 on 5 and 132 DF,  p-value: < 2.2e-16
```

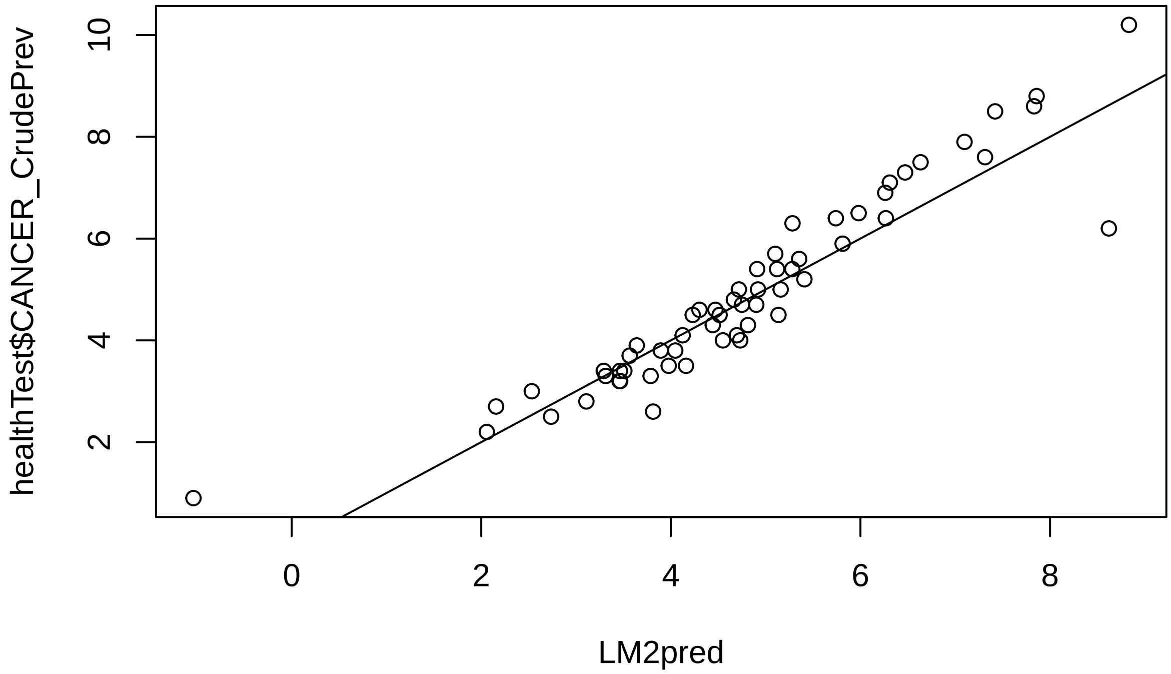
```
par(mfrow=c(2,2))
plot(cancerLM2)
```



```
LM2pred <- predict(cancerLM2, newdata = healthTest)
LM2data<- data.frame(actual=healthTest$CANCER_CrudePrev, predicted= LM2pred)
head(LM2data)
```

```
##    actual predicted
## 2     8.8  7.858934
## 3     2.2  2.058481
## 10    3.3  3.786803
## 11    5.0  5.158567
## 16    0.9 -1.036410
## 19    6.2  8.620857
```

```
par(mfrow=c(1,1))
plot(LM2pred,healthTest$CANCER_CrudePrev)+abline(a=0,b=1)
```



```
## integer(0)
```

```
mae(healthTest$CANCER_CrudePrev,LM2pred)
```

```
## [1] 0.4742241
```

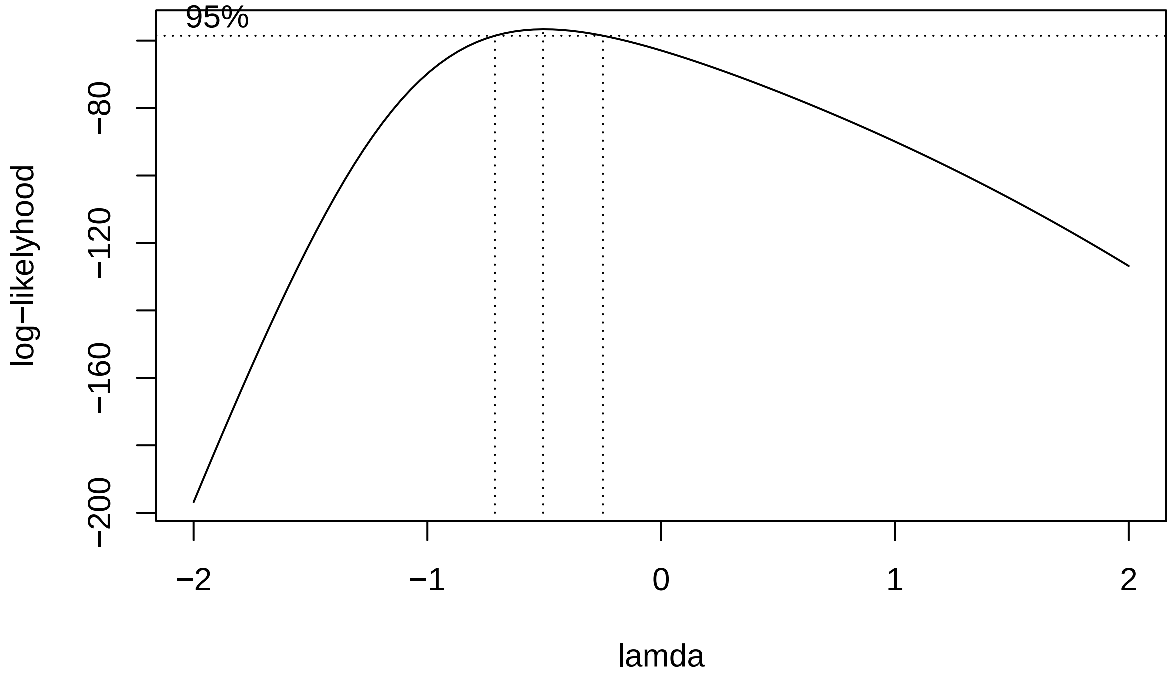
```
mse(healthTest$CANCER_CrudePrev,LM2pred)
```

```
## [1] 0.4318916
```

This multiple linear regression model forced the intercept through zero. As seen in the summary, the model had a residual standard error of 0.61 with an adjusted R squared value of 0.98. Signifying that 98% of the variance in the crude prevalence of cancer can be attributed to crude prevalence of cholesterol screening, chronic obstructive pulmonary disease, blood pressure medication, congenital heart disease, and frequent checkups. The coefficients and model are both significant, with P values less than 0.05. The residuals are evenly centered around zero, and the residuals vs fitted is evenly distributed as well as absent of a pattern.The mean squared error was 0.43 and the mean absolute error was 0.47. The normal qq plot follows the plotted line for the most part. In the predicted vs actual plot using the test set, the 45 degree abline follows the predicted values much closer then that of the single variable linear regression, but there is still some room for improvement.

Transformed Multiple Linear Regression Model

```
boxcox(cancerLM2, lamda = seq(-2,2,1/10), plotit = TRUE, eps =  1/50, xlab = expression(lamda), ylab ="
```

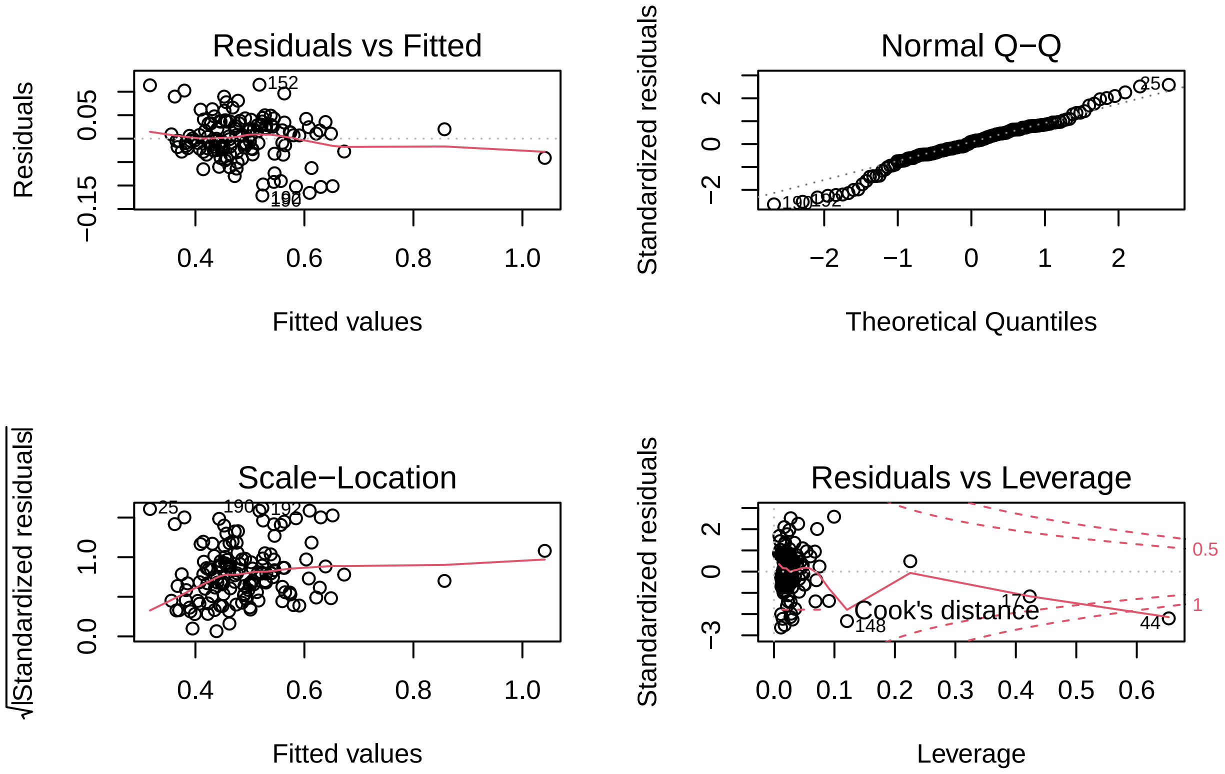


```
healthTrain$boxcox <- (healthTrain$CANCER_CrudePrev)^-0.5
healthTest$boxcox <- (healthTest$CANCER_CrudePrev)^-0.5

cancerLM4<-lm(boxcox ~ COPD_CrudePrev+BPMED_CrudePrev+CHD_CrudePrev+CHOLSCREEN_CrudePrev+CHECKUP_CrudeP
summary(cancerLM4)
```

```
##
## Call:
## lm(formula = boxcox ~ COPD_CrudePrev + BPMED_CrudePrev + CHD_CrudePrev +
##     CHOLSCREEN_CrudePrev + CHECKUP_CrudePrev - 1, data = healthTrain)
##
## Residuals:
##       Min        1Q    Median        3Q       Max
## -0.121274 -0.021267  0.004715  0.030025  0.114943
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## COPD_CrudePrev        0.075577   0.009449   7.998 5.63e-13 ***
## BPMED_CrudePrev      -0.007273   0.002119  -3.432 0.000801 ***
## CHD_CrudePrev        -0.105060   0.013958  -7.527 7.23e-12 ***
## CHOLSCREEN_CrudePrev -0.010457   0.001924  -5.435 2.55e-07 ***
## CHECKUP_CrudePrev     0.030022   0.001810  16.586  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.04634 on 132 degrees of freedom
## Multiple R-squared:  0.9918, Adjusted R-squared:  0.9915
## F-statistic:  3179 on 5 and 132 DF,  p-value: < 2.2e-16
```

```
par(mfrow=c(2,2))
plot(cancerLM4)
```

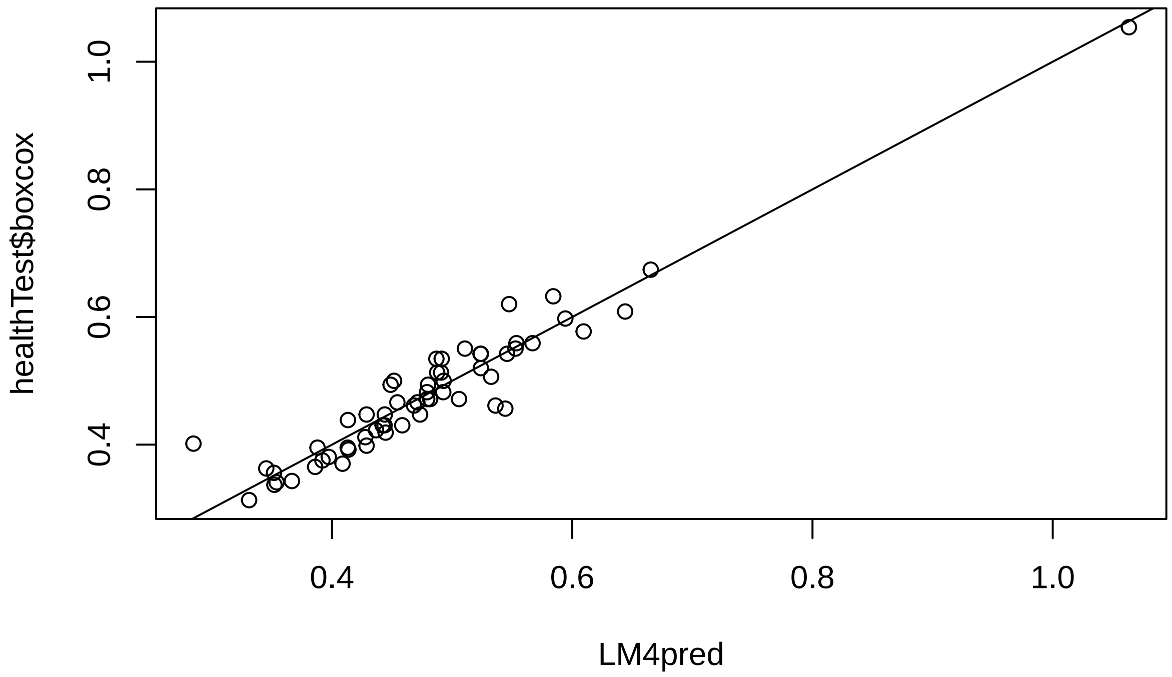


```
par(mfrow=c(1,1))

LM4pred <- predict(cancerLM4, newdata = healthTest)
LM4data<- data.frame(actual=healthTest$boxcox, predicted= LM4pred)
head(LM4data)
```

```
##       actual predicted
## 2  0.3370999 0.3520179
## 3  0.6741999 0.6653534
## 10 0.5504819 0.5106370
## 11 0.4472136 0.4287716
## 16 1.0540926 1.0634475
## 19 0.4016097 0.2846090
```

```
par(mfrow=c(1,1))
plot(LM4pred, healthTest$boxcox)+abline(a=0,b=1)
```



```
## integer(0)
```

```
mae(healthTest$boxcox,LM4pred)
```

```
## [1] 0.02374746
```

```
mse(healthTest$boxcox,LM4pred)
```

```
## [1] 0.001041487
```

This multiple linear regression model forced the intercept through zero, contained the same explanatory variables as tested in the multiple linear regression model previously, and contained a transformation of cancer crude prevalence that was found using a boxcox analysis. As seen in the summary, the model had a residual standard error of 0.04 with an adjusted R squared value of 0.99. Signifying that 99% of the variance in the crude prevalence of cancer can be attributed to crude prevalence of cholesterol screening, chronic obstructive pulmonary disease, blood pressure medication, congenital heart disease, and frequent checkups. The coefficients and model are both significant, with P values less than 0.05. The residuals are evenly centered around zero, and the residuals vs fitted is evenly distributed as well as absent of a pattern.The mean squared error was 0.001 and the mean absolute error was 0.023. The normal qq plot follows the plotted line for the most part. In the predicted vs actual plot using the test set, the 45 degree abline follows the predicted values much closer then that of the single variable linear regression and the previous non transformed multiple linear regression.

Support Vector Machine Classification Model

```
healthTrain$new[healthTrain$CANCER_CrudePrev <= 1] <- 1
healthTest$new[healthTest$CANCER_CrudePrev <= 1] <- 1

healthTrain$new[healthTrain$CANCER_CrudePrev > 1] <- 2
healthTest$new[healthTest$CANCER_CrudePrev > 1] <- 2

healthTrain$new[healthTrain$CANCER_CrudePrev > 2] <- 3
healthTest$new[healthTest$CANCER_CrudePrev > 2] <- 3

healthTrain$new[healthTrain$CANCER_CrudePrev > 3] <- 4
healthTest$new[healthTest$CANCER_CrudePrev > 3] <- 4

healthTrain$new[healthTrain$CANCER_CrudePrev > 4] <- 5
healthTest$new[healthTest$CANCER_CrudePrev > 4] <- 5

healthTrain$new[healthTrain$CANCER_CrudePrev > 6] <- 7
healthTest$new[healthTest$CANCER_CrudePrev > 6] <- 7

healthTrain$new[healthTrain$CANCER_CrudePrev > 7] <- 8
healthTest$new[healthTest$CANCER_CrudePrev > 7] <- 8

healthTrain$new[healthTrain$CANCER_CrudePrev > 8] <- 9
healthTest$new[healthTest$CANCER_CrudePrev > 8] <- 9

healthTrain$new[healthTrain$CANCER_CrudePrev > 9] <- 10
healthTest$new[healthTest$CANCER_CrudePrev > 9] <- 10

head(healthTrain$CANCER_CrudePrev)
```

```
## [1] 2.3 2.4 2.5 2.4 5.8 4.3
```

```
head(healthTest$CANCER_CrudePrev)
```

```
## [1] 8.8 2.2 3.3 5.0 0.9 6.2
```

```
head(healthTrain$new)
```

```
## [1] 3 3 3 3 5 5
```

```
head(healthTest$new)
```

```
## [1] 9 3 4 5 1 7
```

As discussed in the section about data cleaning, some transformations were required in order to produce an accurate classification model for cancer crude prevalence. Before transformation the crude prevalence data was down to the tenths place of a decimal. To make classification a more viable solution, each value was rounded up to the nearest whole number. This created a new variable with 10 distinct possibilities, then converted to a factor variable for the support vector machine. In terms of classification models, the support vector machine model utilizing the linear kernel resulted in a greater accuracy percentage for classifications of the test set when compared to naive bayes, principle component analysis, and support vector machines using the polynomial or radial kernel functions.

```
set.seed(123)
model.sv1 <- svm(as.factor(new)~ CHOLSCREEN_CrudePrev + CHD_CrudePrev, data = healthTrain, kernel = "li
summary(model.sv1)
```

```
##
## Call:
## svm(formula = as.factor(new) ~ CHOLSCREEN_CrudePrev + CHD_CrudePrev,
##     data = healthTrain, kernel = "linear")
##
##
## Parameters:
##    SVM-Type:  C-classification
##  SVM-Kernel:  linear
##        cost:  1
##
## Number of Support Vectors:  95
##
##  ( 12 30 14 28 6 1 3 1 )
##
##
## Number of Classes:  8
##
## Levels:
##  1 2 3 4 5 7 8 9
```

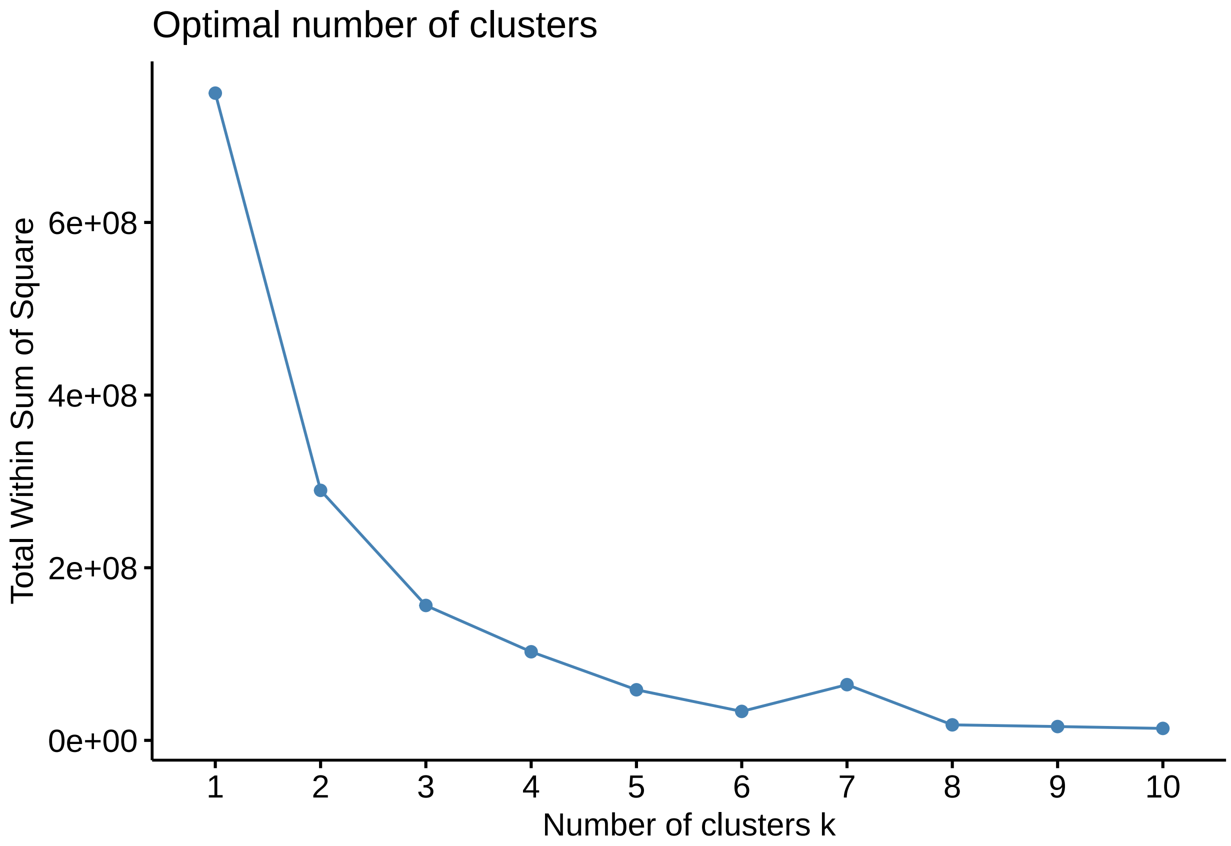
```
table(predict(model.sv1, healthTest[(1:length(model.sv1))]) == healthTest$new)/length(healthTest$new)
```

```
##
##     FALSE      TRUE
## 0.3050847 0.6949153
```

From the summary command we can see that the support vector machine model contains a total of 95 support vectors with 8 different classes. When testing the classification accuracy on the test set, the model accurately classified 69.5% of all observations into the correct class out of 8, using only cholesterol screening and congenital heart disease as explanatory variables.

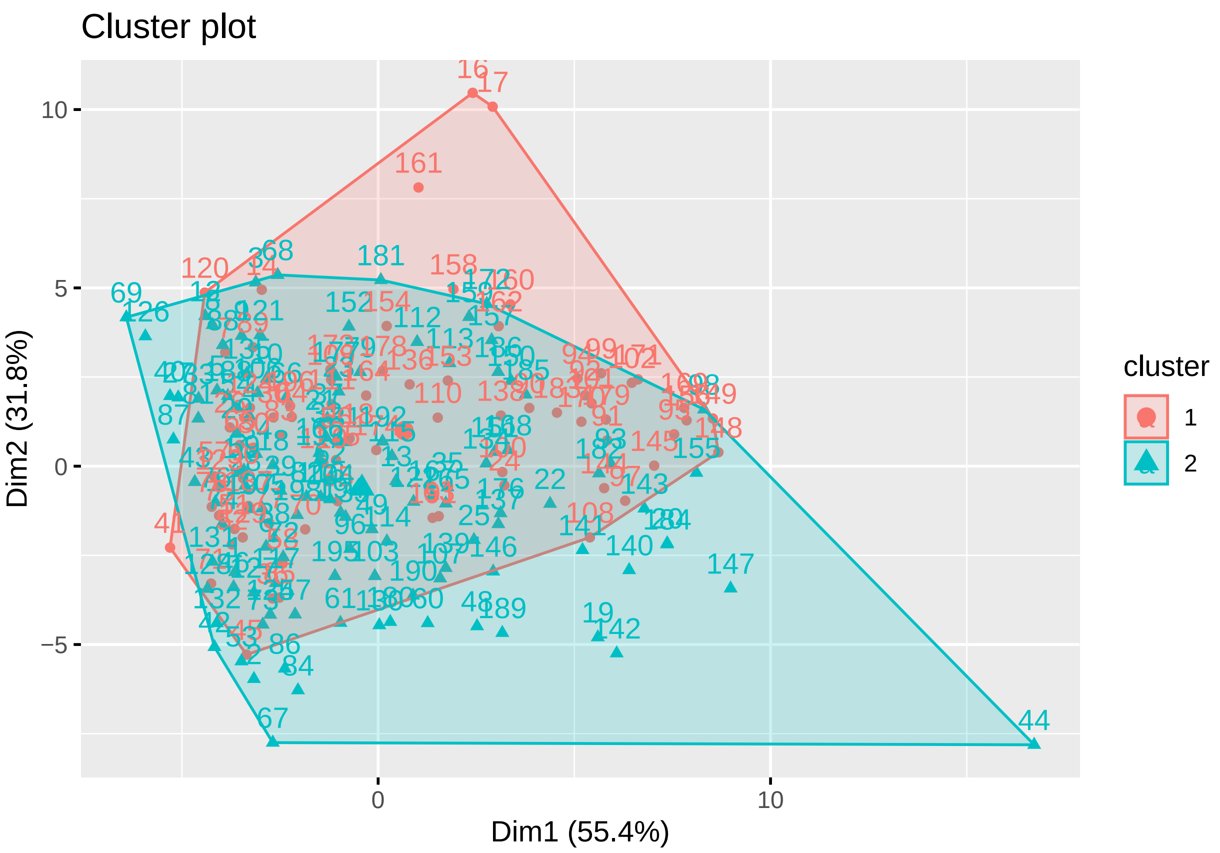
Non Parametric K Means Clustering Model

```r
fviz_nbclust(healthnew, kmeans, method="wss")
```

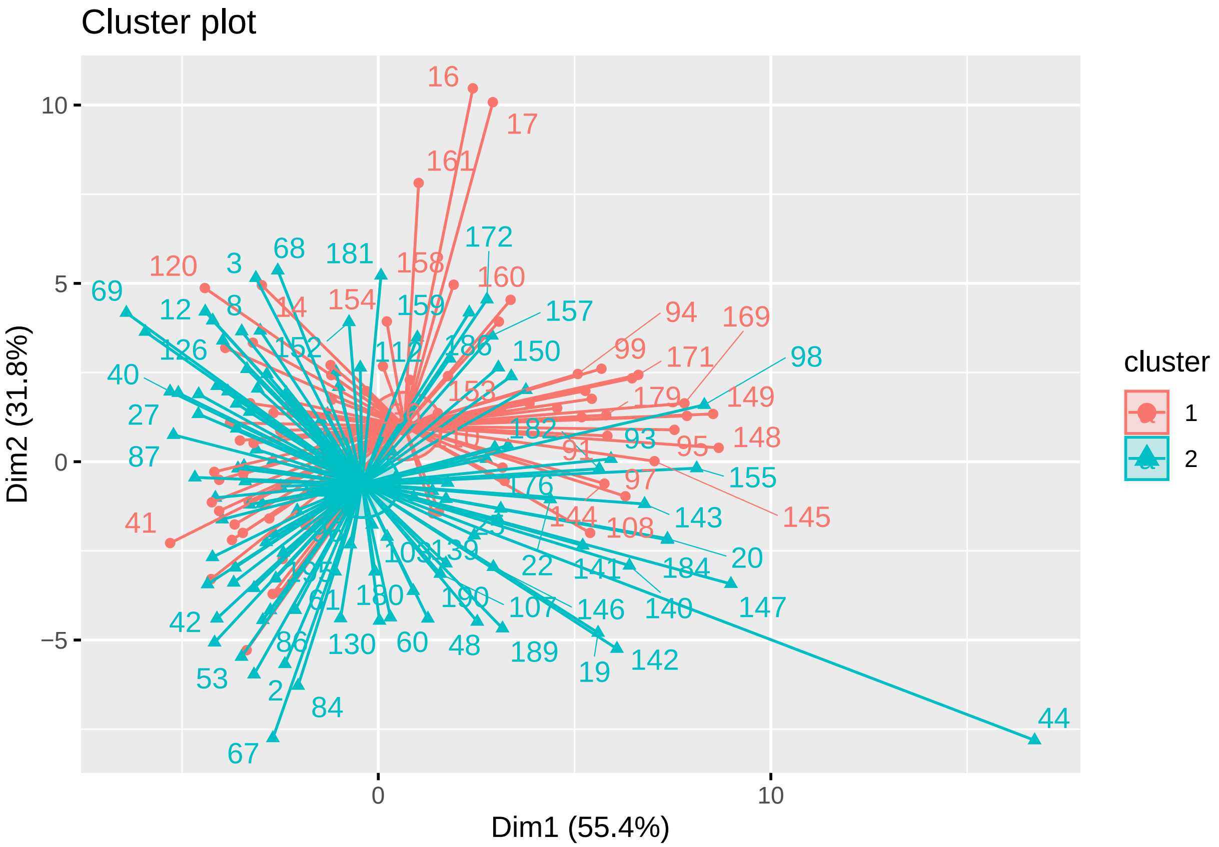


Examining the WSS elbow plot, there isn't a definite elbow and it more so resembles a negative exponential curve. Multiple models will be created and compared with k values equal to 2 and 4.

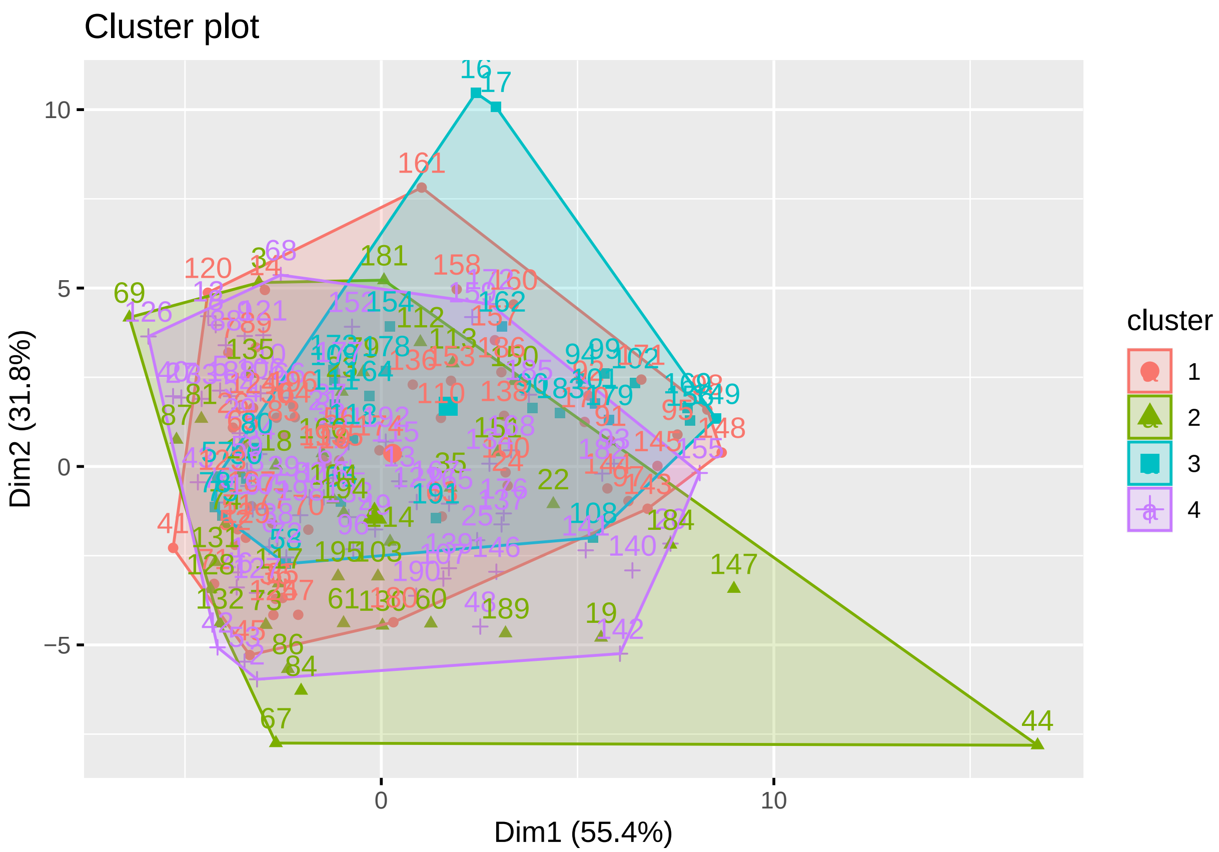
```r
km2 <- kmeans(healthnew, 2)
fviz_cluster(km2, data=healthnew)
```



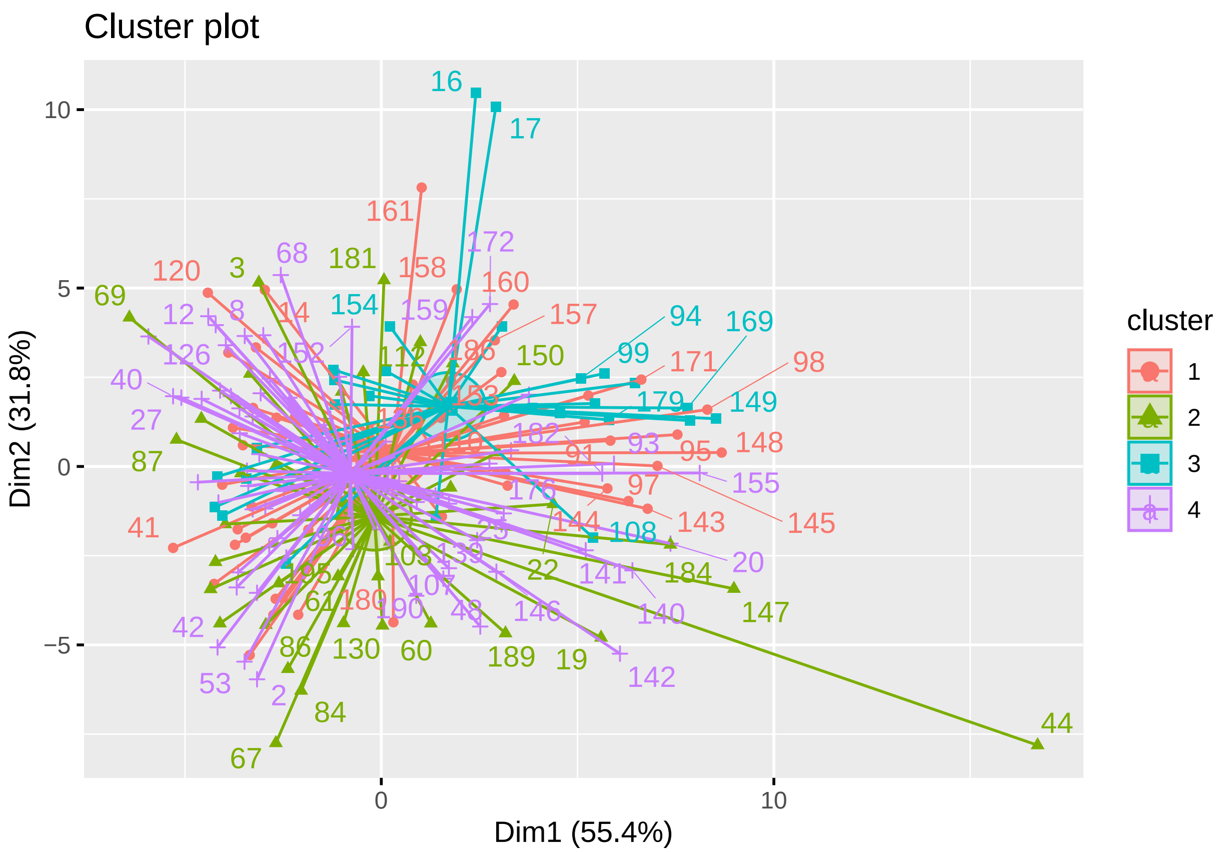
```r
fviz_cluster(km2, data = healthnew,
             ellipse.type = "euclid",
             star.plot = T,
             repel = T,
             ggtheme = theme())
```



```
km4 <- kmeans(healthnew, 4)
fviz_cluster(km4, data=healthnew)
```

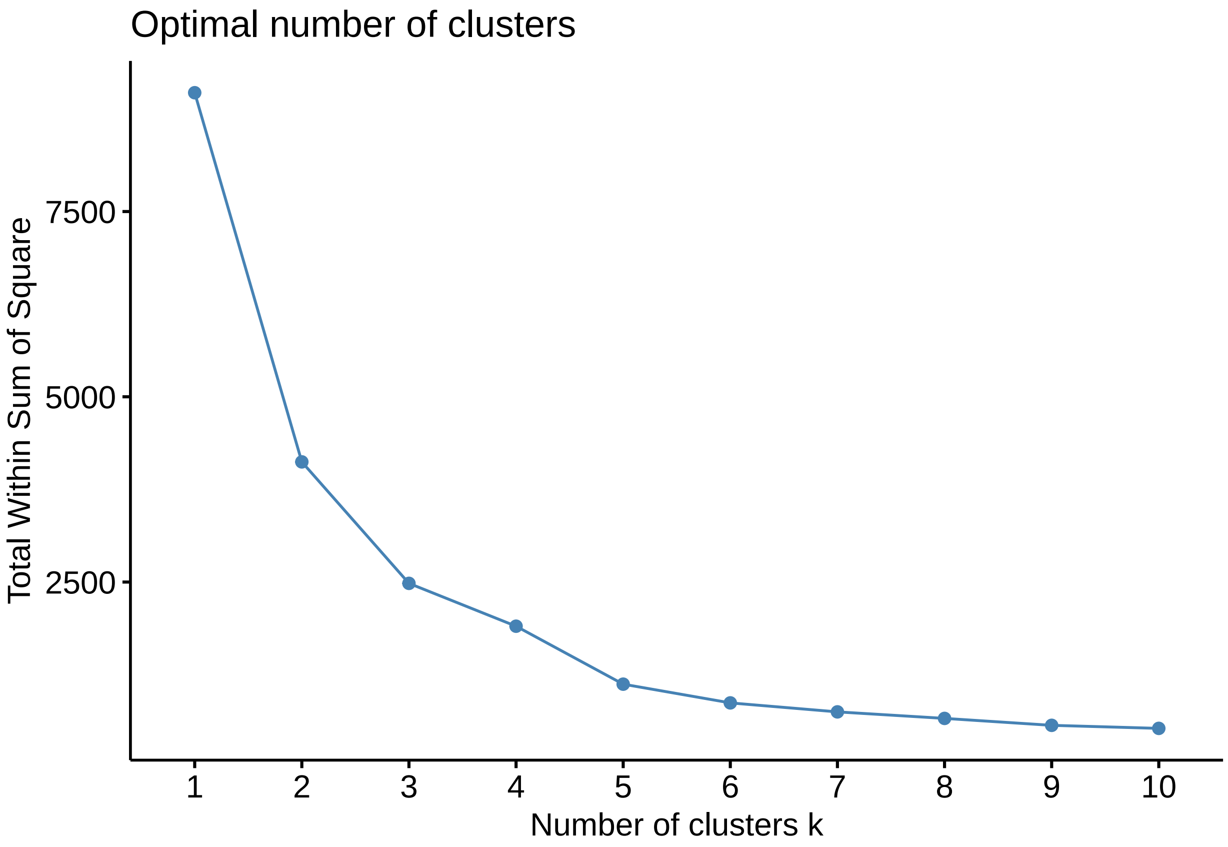


```r
fviz_cluster(km4, data = healthnew,
             ellipse.type = "euclid",
             star.plot = T,
             repel = T,
             ggtheme = theme())
```

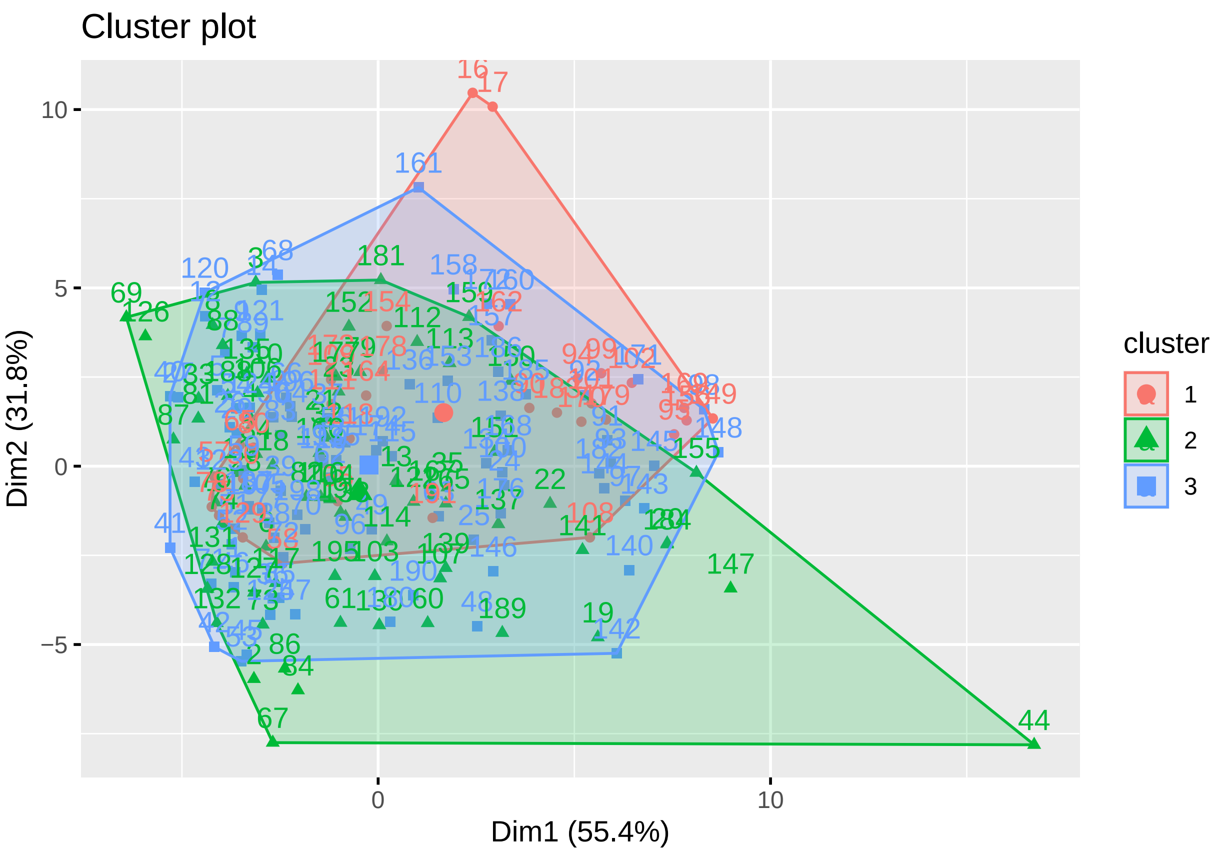


As evidence from the within cluster sum of squares and the immense overlap in the cluster plots, a non parametric k means clustering approach to clustering the data is not suitable for any value of k. Using the data set that contains all variables of crude prevalence may not be the best approach for clustering analysis. A new data frame will be constructed consisting of cancer crude prevalence, cholesterol crude prevalence, congenital heart disease crude prevalence and a k means clustering analysis will be attempted again using variables that are previously shown to contain correlation.

```
healthclust<- data.frame(healthnew$CANCER_CrudePrev,healthnew$CHOLSCREEN_CrudePrev, healthnew$CHD_Crude
fviz_nbclust(healthclust, kmeans, method="wss")
```

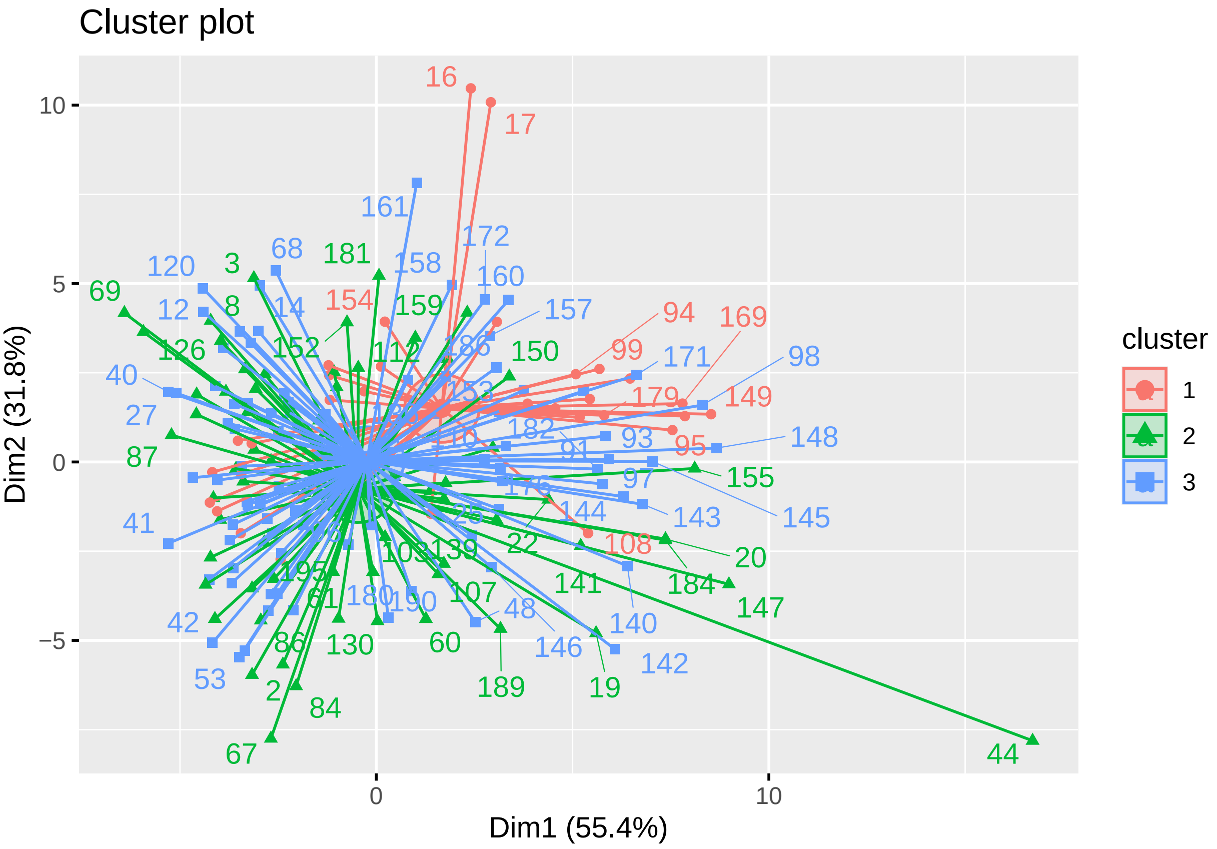


```
km3 <- kmeans(healthnew, 3)
fviz_cluster(km3, data=healthnew)
```



```
fviz_cluster(km3, data = healthnew,
             ellipse.type = "euclid",
             star.plot = T,
             repel = T,
             ggtheme = theme())
```

```
## Warning: ggrepel: 119 unlabeled data points (too many overlaps). Consider
## increasing max.overlaps
```



Once again, the within cluster sum of squares is rather large and the cluster plot contains immense overlap deeming it not suitable for any further analysis.

Conclusion

To conclude the finding of the study, the multiple linear regression model containing the boxcox transformation and the support vector machine were most useful in accurately predicting and classifying observations with cancer crude prevalence as the prediction variable. The most significant explanatory variables were crude prevalence of cholesterol screening, chronic obstructive pulmonary disease, blood pressure medication, congenital heart disease, and frequent checkups. Examining coefficient estimates from the transformed multiple linear regression model, it can be stated that for a one unit increase in crude prevalence of chronic obstructive pulmonary disease, a 0.07 increase in cancer crude prevalence can be expected. For a one unit increase in crude prevalence of blood pressure medication, a 0.007 decrease in cancer crude prevalence can be expected. For a one unit increase in crude prevalence of congenital heart disease, a 0.11 decrease in cancer crude prevalence can be expected. For a one unit increase in crude prevalence of cholesterol screening, a 0.01 decrease in cancer crude prevalence can be expected. Finally, for a one unit increase in crude prevalence of routine checkups, a 0.03 increase in cancer crude prevalence can be expected. These coefficient estimates give the most insight into potential solutions for the mayor of cityX if he or she so wishes to decrease the crude prevalence of cancer.