Dimensionality Reduction

Load and Clean the data

Dataset: Kansas City House Data via Kaggle.

```
df <- read.csv("cubic_zirconia.csv")
df$X <- NULL

df$cut <- as.numeric(factor(df$cut, levels = c("Fair", "Good", "Very Good", "Premium", "Ideal")))
df$color <- as.numeric(as.factor(df$color))
df$clarity <- as.numeric(factor(df$clarity, levels = c("I1", "SI2", "SI1", "VS2", "VS1", "VVS2", "VVS1"

df <- df[!(df$x == 0),]
df <- df[!(df$y == 0),]
df <- df[!(df$z == 0),]

set.seed(1234)
i <- sample(1:nrow(df), 0.8*nrow(df), replace=FALSE)
train <- df[i,]
test <- df[-i,]
```

Explore the Data

Data is further explored statistically and graphically for PCA and LDA since these two algorithms differ in their mathematical approaches.

```
str(train)
```

```
## 'data.frame':    21566 obs. of  10 variables:
##  $ carat  : num  0.39 0.28 1.45 0.9 0.3 0.71 1 1.18 0.34 2.01 ...
##  $ cut    : num  5 2 1 2 5 2 2 5 5 2 ...
##  $ color  : num  5 2 3 1 4 2 7 4 6 6 ...
##  $ clarity: num  6 7 2 2 8 2 4 5 3 2 ...
##  $ depth  : num  62 64.6 64.4 60.3 62.6 57.8 58.7 61.9 61.9 56.9 ...
##  $ table  : num  55 55 58 64 54 60 61 56 56 59 ...
##  $ x      : num  4.68 4.15 7.17 6.1 4.3 5.86 6.5 6.81 4.47 8.41 ...
##  $ y      : num  4.7 4.18 7.11 6.14 4.33 5.83 6.52 6.76 4.5 8.3 ...
##  $ z      : num  2.91 2.69 4.59 3.69 2.7 3.38 3.82 4.2 2.77 4.74 ...
##  $ price  : int  794 492 6455 3534 826 2215 3769 8556 469 11312 ...
```

Principle Component Analysis - PCA

```
library("factoextra")
```

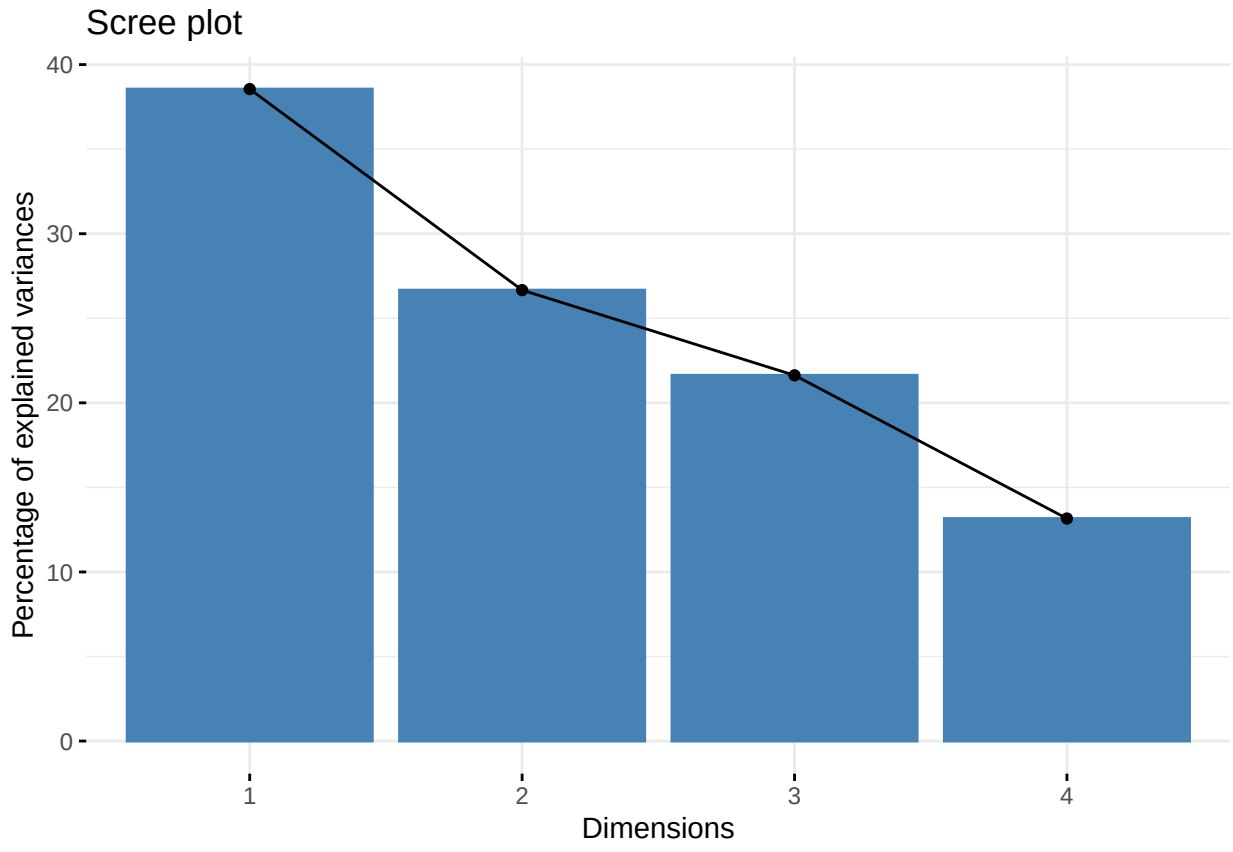
```
## Loading required package: ggplot2
```

```
## Welcome! Want to learn more? See two factoextra-related books at https://goo.gl/ve3WBa
```

```
pca1 <- prcomp(train[,1:4], scale = TRUE)
summary(pca1)
```

```
## Importance of components:
##                           PC1    PC2    PC3    PC4
## Standard deviation     1.2418 1.0328 0.9301 0.7255
## Proportion of Variance 0.3855 0.2666 0.2163 0.1316
## Cumulative Proportion  0.3855 0.6522 0.8684 1.0000
```

```
#Visualize eigenvalues (scree plot)
fviz_eig(pca1)
```



Linear Discriminant Analysis (LDA)

```
library(MASS)
lda1 <- lda(price~., data = train)
head(lda1$means)
```

```
##     carat cut color clarity depth table     x    y     z
## 326  0.22 4.5   2.0     2.5 60.65    58 3.920 3.91 2.370
## 335  0.31 2.0   7.0     2.0 63.30    58 4.340 4.35 2.750
## 336  0.24 3.0   6.5     6.5 62.55    57 3.945 3.97 2.475
## 338  0.23 3.0   5.0     5.0 59.40    61 4.000 4.05 2.390
## 345  0.32 4.0   2.0     1.0 60.90    58 4.380 4.42 2.680
## 348  0.30 5.0   6.0     2.0 62.00    54 4.310 4.34 2.680
```

Accuracy Loss

There is possible accuracy when applying either PCA or LDA because the algorithms will not take into the account actual target variable when choosing which features to reduce. These algorithms could deem features with high variance as important features, but such features may not even have anything to do with the prediction target.Additionally, PCA and LDA are both very sensitive to outliers, which can lead to misleading conclusions when outliers are present. Hence, it is important to perform proper and thorough data preprocessing.