---
title: "Similarity Part 4: PCA and LDA"
output: html_notebook

Made by: Jonathan Blade

Data set used can be found here https://www.kaggle.com/datasets/teejmahal20/airline-passenger-satisfaction?resource=download

(Note: The original data set was divided into two files. I used the file "train.csv" and took a train and test sample from said file.)

This data set is used for classification where the goal is to determine whether a customer was satisfied with their flight.

```{r}

library(caret)

planeData <- read.csv("C:\\Users\\18327\\Desktop\\Academics\\Academics Fall 2022\\Machine Learning\\Portfolio Similarity\\train.csv", header=TRUE)

planeData$satisfaction <- as.factor(planeData$satisfaction)


sapply(planeData, function(x) sum(is.na(x)))

planeData <- planeData[(complete.cases(planeData)),]
sum(is.na(planeData))

planeData <- planeData[(complete.cases(planeData)),]
sum(is.na(planeData))

set.seed(12345)

sample <- sample(c(TRUE,FALSE), nrow(planeData), replace=TRUE, prob=c(0.8,0.2))
train <- planeData[sample, ]
test <- planeData[!sample, ]

summary(train)

pca_out <- preProcess(train[,1:24], method=c("center", "scale", "pca"))

pca_out

train_pca <- predict(pca_out, train[,1:24])
test_pca <- predict(pca_out, test[,])

train_df <- data.frame(train_pca$PC1, train_pca$PC2, train_pca$PC3, train_pca$PC4, train_pca$PC5, train_pca$PC6, train_pca$PC7, train_pca$PC8, train_pca$PC9, train_pca$PC10, train_pca$PC11, train_pca$PC12, train_pca$PC13, train_pca$PC14, train_pca$PC15, train_pca$PC16, train$satisfaction)
```

```
test_df <- data.frame(test_pca$PC1, test_pca$PC2, test_pca$PC3, test_pca$PC4,
test_pca$PC5, test_pca$PC6, test_pca$PC7, test_pca$PC8, test_pca$PC9,
test_pca$PC10, test_pca$PC11, test_pca$PC12, test_pca$PC13, test_pca$PC14,
test_pca$PC15, test_pca$PC16, test$satisfaction)

library(class)

set.seed(12345)

pred <- knn(train=train_df[,1:16], test=test_df[,1:16], cl=train_df[,17], k=3)
meanPCA <- mean(pred == test$satisfaction)
print(paste("Accuracy with PCA: ", meanPCA))

library(tree)
colnames(train_df) <- c("PC1", "PC2", "PC3", "PC4", "PC5", "PC6", "PC7",
"PC8", "PC9", "PC10", "PC11", "PC12", "PC13", "PC14", "PC15", "PC16",
"Satisfaction")
colnames(test_df) <- c("PC1", "PC2", "PC3", "PC4", "PC5", "PC6", "PC7", "PC8",
"PC9", "PC10", "PC11", "PC12", "PC13", "PC14", "PC15", "PC16", "Satisfaction")

set.seed(12345)

tree1 <- tree(Satisfaction~., data = train_df)
plot(tree1)
text(tree1, cex=0.5, pretty=0)

pred2 <- predict(tree1, newdata=test_df, type="class")
pcaTree <- mean(pred2==test$satisfaction)
print(paste("PCA with tree: ", pcaTree))

```

The classification with PCA is slightly less accurate than without it.
However, PCA's main function is to reduce the number of dimensions of the
data.
A consequence of this reduction of dimensions is a loss of interpretability.
This can be seen in the above tree diagram, typically a plot that enhances
interpretability, is now more difficult to understand.

```{R}
library(MASS)

lda1 <- lda(satisfaction~., data=train)
lda1$means

lda_pred <- predict(lda1, newdata = test, type="class")

meanLDA <- mean(lda_pred$class==test$satisfaction)
print(paste("Accuracy with LDA: ", meanLDA))
```
With LDA, the accuracy is lower than PCA but the interpretability is
maintained.