Dimensionality Reduction

Khang Thai, David Park, Jonathan Ho, David Favela

Reading in the Dataset

```
library(caret)
```

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

```
## Loading required package: lattice
```

```
Invistico_Airline <- read.csv("Invistico_Airline.csv", header=TRUE)
str(Invistico_Airline)
```

```
## 'data.frame':    129880 obs. of  23 variables:
##  $ satisfaction                     : chr  "satisfied" "satisfied" "satisfied" "satisfied"
...
##  $ Gender                           : chr  "Female" "Male" "Female" "Female" ...
##  $ Customer.Type                    : chr  "Loyal Customer" "Loyal Customer" "Loyal Customer"
"Loyal Customer" ...
##  $ Age                              : int  65 47 15 60 70 30 66 10 56 22 ...
##  $ Type.of.Travel                   : chr  "Personal Travel" "Personal Travel" "Personal Trav
el" "Personal Travel" ...
##  $ Class                            : chr  "Eco" "Business" "Eco" "Eco" ...
##  $ Flight.Distance                  : int  265 2464 2138 623 354 1894 227 1812 73 1556 ...
##  $ Seat.comfort                     : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Departure.Arrival.time.convenient: int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Food.and.drink                   : int  0 0 0 0 0 0 0 0 0 0 ...
##  $ Gate.location                    : int  2 3 3 3 3 3 3 3 3 3 ...
##  $ Inflight.wifi.service            : int  2 0 2 3 4 2 2 2 5 2 ...
##  $ Inflight.entertainment           : int  4 2 0 4 3 0 5 0 3 0 ...
##  $ Online.support                   : int  2 2 2 3 4 2 5 2 5 2 ...
##  $ Ease.of.Online.booking           : int  3 3 2 1 2 2 5 2 4 2 ...
##  $ On.board.service                 : int  3 4 3 1 2 5 5 3 4 2 ...
##  $ Leg.room.service                 : int  0 4 3 0 0 4 0 3 0 4 ...
##  $ Baggage.handling                 : int  3 4 4 1 2 5 5 4 1 5 ...
##  $ Checkin.service                  : int  5 2 4 4 4 5 5 5 5 3 ...
##  $ Cleanliness                      : int  3 3 4 1 2 4 5 4 4 4 ...
##  $ Online.boarding                  : int  2 2 2 3 5 2 3 2 4 2 ...
##  $ Departure.Delay.in.Minutes       : int  0 310 0 0 0 0 17 0 0 30 ...
##  $ Arrival.Delay.in.Minutes         : int  0 305 0 0 0 0 15 0 0 26 ...
```

Convert satisfaction into a factor

```
Invistico_Airline$satisfaction <- as.factor(Invistico_Airline$satisfaction)
```

Create New Columns Rating Sum & Rating Mean

```
Invistico_Airline$ratingSum <- as.numeric(apply(Invistico_Airline[,8:21], 1, sum))
Invistico_Airline$ratingMean <- c(Invistico_Airline$ratingSum/14)
```

Create train & test sets

```
i <- sample(1:nrow(Invistico_Airline), 0.8*nrow(Invistico_Airline), replace = FALSE)
trainAirline <- Invistico_Airline[i,]
testAirline <- Invistico_Airline[-i,]
set.seed(3)
```

Clean out columns not needed

```
trainAirline <- trainAirline[,c(4,7,25,1)]
testAirline <- testAirline[,c(4,7,25,1)]
```

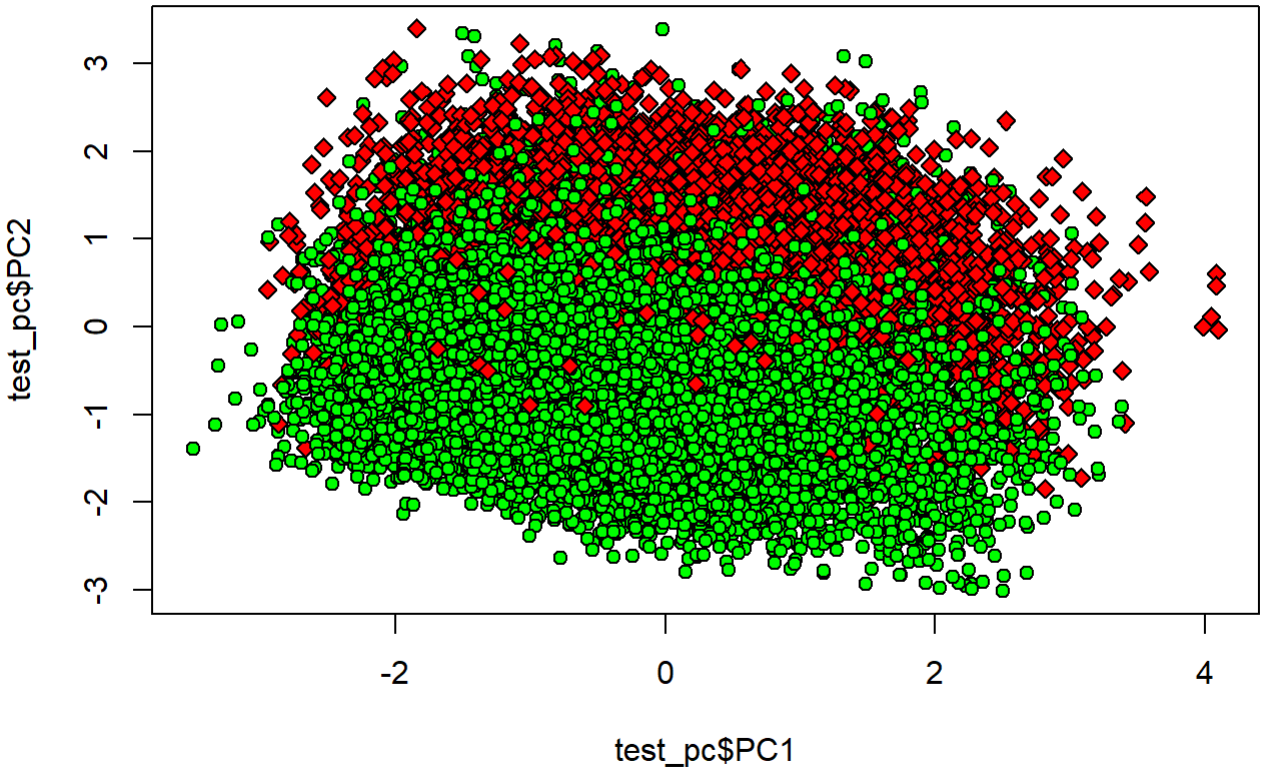
PCA

```
pca_out <- preProcess(trainAirline[,1:3], method=c("center", "scale", "pca"))
pca_out
```

```
## Created from 103904 samples and 3 variables
##
## Pre-processing:
##   - centered (3)
##   - ignored (0)
##   - principal component signal extraction (3)
##   - scaled (3)
##
## PCA needed 3 components to capture 95 percent of the variance
```

Plotting the PCA with the Test Data

```
train_pc <- predict(pca_out, trainAirline[,1:3])
test_pc <- predict(pca_out, testAirline[,])
plot(test_pc$PC1, test_pc$PC2, pch=c(23,21,22)[unclass(test_pc$satisfaction)], bg=c("red", "gree
n", "blue")[unclass(testAirline$satisfaction)])
```



##

Finding the accuracy of PCA

```
train_df <- data.frame(train_pc$PC1, train_pc$PC2,  trainAirline$satisfaction)
test_df <- data.frame(test_pc$PC1, test_pc$PC2, testAirline$satisfaction)
library(class)

pred <- knn(train = train_df[,1:2], test = test_df[,1:2], cl=train_df[,3], k=3)
mean(pred==testAirline$satisfaction)
```

```
## [1] 0.684863
```

Regression comparison to PCA

```
glm1 <- glm(satisfaction~., data=trainAirline, family=binomial)
```

```
probs <- predict(glm1, newdata = testAirline, type="response")
pred <- ifelse(probs>0.5,2,1)
acc1 <- mean(pred==as.integer(testAirline$satisfaction))
print(paste("glm1 accuracy = ", acc1))
```

```
## [1] "glm1 accuracy =  0.735178626424392"
```

Results

The amount of accuracy lost from Regression to PCA was about 5% accuracy. Regression had a 73% accuracy while PCA had a 68% accuracy.

LDA

```
library(MASS)
lda1 <- lda(satisfaction~., data=trainAirline)
lda1$means
```

```
##                   Age Flight.Distance ratingMean
## dissatisfied 37.45213        2027.184   2.935110
## satisfied    41.06625        1946.415   3.617817
```

Predicting the satisfaction based on the test data

```
lda_pred <- predict(lda1, newdata=testAirline, type="class")
head(lda_pred$class)
```

```
## [1] dissatisfied dissatisfied dissatisfied dissatisfied dissatisfied
## [6] dissatisfied
## Levels: dissatisfied satisfied
```

Regression Comparison to LDA

```
glm1 <- glm(satisfaction~., data=trainAirline, family=binomial)
```

```
probs <- predict(glm1, newdata = testAirline, type="response")
pred <- ifelse(probs>0.5,2,1)
acc1 <- mean(pred==as.integer(testAirline$satisfaction))
print(paste("glm1 accuracy = ", acc1))
```

```
## [1] "glm1 accuracy =  0.735178626424392"
```

```
table(pred, as.integer(testAirline$satisfaction))
```

```
##
## pred     1     2
##    1  7829  2988
##    2  3891 11268
```

```
mean(lda_pred$class==testAirline$satisfaction)
```

```
## [1] 0.7352556
```

Results

Here, the accuracy for LDA was 73.44% while the accuracy of regression was 73.49% so the amount of accuracy lost was only .05%.