STT 461 ICA 9

Derien Weatherspoon

2023-04-01

For the problems below, the dataset found in OJ.csv. The target is Purchase.

```
df <- read.csv("OJ.csv")
library(tree)
```

Take a look at the different fields. Are there fields which can be calculated from other fields?

```
str(df)
```

```
## 'data.frame':    1070 obs. of  20 variables:
##  $ X             : int  1 2 3 4 5 6 7 8 9 10 ...
##  $ Purchase      : chr  "CH" "CH" "CH" "MM" ...
##  $ WeekofPurchase: int  237 239 245 227 228 230 232 234 235 238 ...
##  $ StoreID       : int  1 1 1 1 7 7 7 7 7 7 ...
##  $ PriceCH       : num  1.75 1.75 1.86 1.69 1.69 1.69 1.69 1.75 1.75 1.75 ...
##  $ PriceMM       : num  1.99 1.99 2.09 1.69 1.69 1.99 1.99 1.99 1.99 1.99 ...
##  $ DiscCH        : num  0 0 0.17 0 0 0 0 0 0 0 ...
##  $ DiscMM        : num  0 0.3 0 0 0 0 0.4 0.4 0.4 0.4 ...
##  $ SpecialCH     : int  0 0 0 0 0 0 1 1 0 0 ...
##  $ SpecialMM     : int  0 1 0 0 0 1 1 0 0 0 ...
##  $ LoyalCH       : num  0.5 0.6 0.68 0.4 0.957 ...
##  $ SalePriceMM   : num  1.99 1.69 2.09 1.69 1.69 1.99 1.59 1.59 1.59 1.59 ...
##  $ SalePriceCH   : num  1.75 1.75 1.69 1.69 1.69 1.69 1.69 1.75 1.75 1.75 ...
##  $ PriceDiff     : num  0.24 -0.06 0.4 0 0 0.3 -0.1 -0.16 -0.16 -0.16 ...
##  $ Store7        : chr  "No" "No" "No" "No" ...
##  $ PctDiscMM     : num  0 0.151 0 0 0 ...
##  $ PctDiscCH     : num  0 0 0.0914 0 0 ...
##  $ ListPriceDiff : num  0.24 0.24 0.23 0 0 0.3 0.3 0.24 0.24 0.24 ...
##  $ STORE         : int  1 1 1 1 0 0 0 0 0 0 ...
##  $ y             : int  0 0 0 1 0 0 0 0 0 0 ...
```

It looks like there are some fields which can be calculated from others, like PriceDiff and ListPriceDiff.

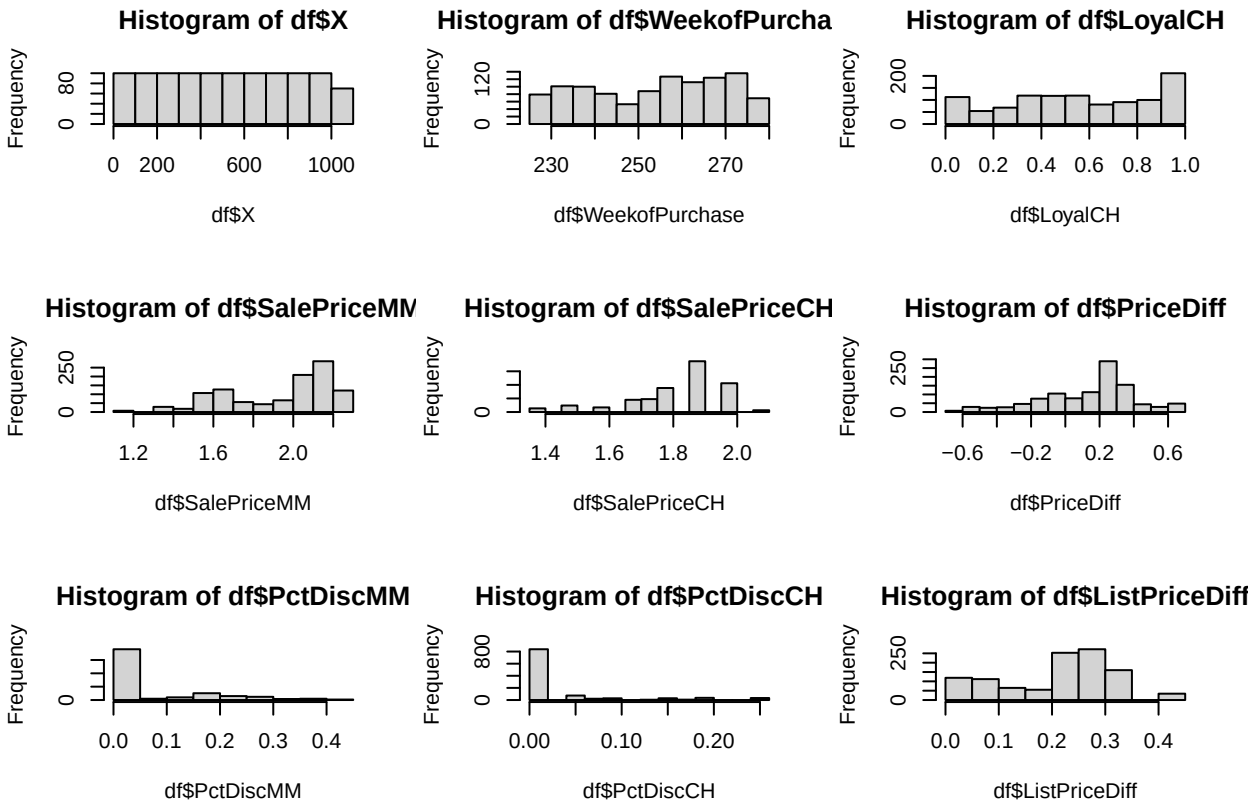
Is StoreID a numeric field? Should it be?

StoreID is a integer here, I think transforming it to a numeric would work better.

```
StoreID_num <- as.numeric(df$StoreID)
```

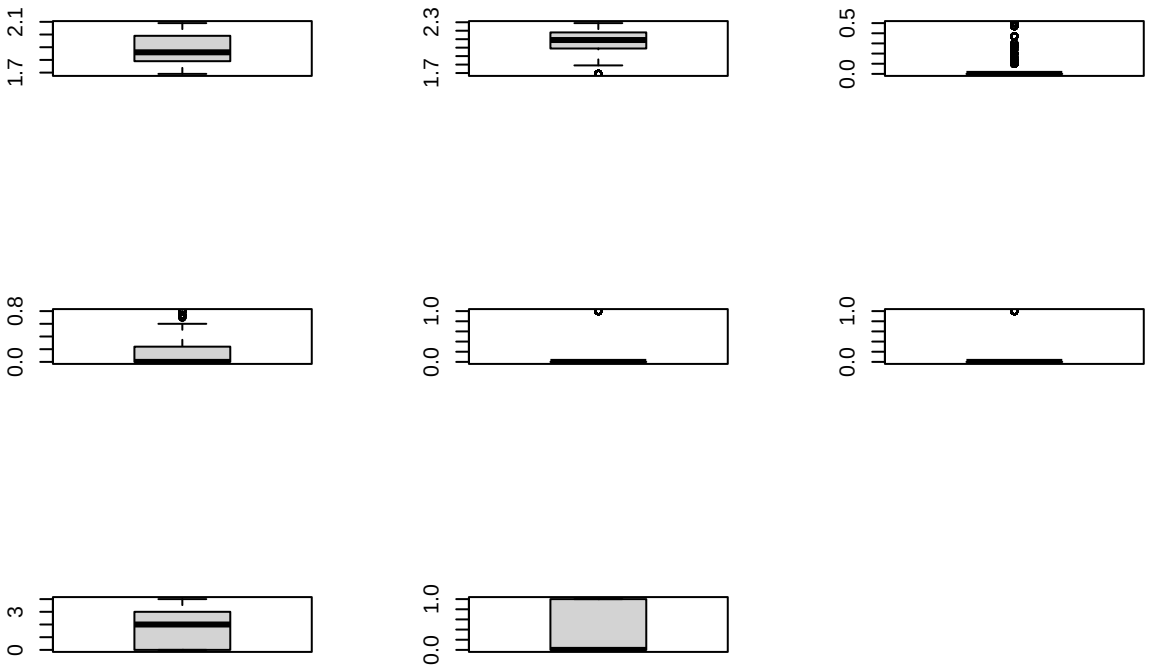
Do some exploratory data analysis to see how different fields relate to Purchase (mainly the histograms as done in the 1st video). Which fields seem relevant?

```
# For the numeric variables

par(mfrow=c(3,3))
hist(df$X)
hist(df$WeekofPurchase)
hist(df$LoyalCH)
hist(df$SalePriceMM)
hist(df$SalePriceCH)
hist(df$PriceDiff)
hist(df$PctDiscMM)
hist(df$PctDiscCH)
hist(df$ListPriceDiff)
```




```
# For the categorical variables

par(mfrow=c(3,3))
boxplot(df$PriceCH)
boxplot(df$PriceMM)
```

```
boxplot(df$DiscCH)
boxplot(df$DiscMM)
boxplot(df$SpecialCH)
boxplot(df$SpecialMM)
boxplot(df$STORE)
boxplot(df$y)
```



Create a 75/25 train/test split on the data.

```
rows <- 1:nrow(df)
train_split <- sample(rows, 0.75*length(rows))
test_split <- rows[-train_split]
train <- df[train_split,]
test <- df[test_split,]
```

Create a logistic regression model with Purchase and 2-3 of the variables as inputs. How does the confusion matrix look for the test dataset?

```
#glm.fit <- glm(Purchase ~ SpecialCH + SpecialMM + DiscCH + DiscMM,  data = test, family = "binomial")
#error?
```

Create a decision tree model with Purchase and 2-3 of the variables as inputs. How does the confusion matrix look for the test dataset?

```
#confusionMatrix(model, df$purchase)
# can't do confusion matrix without model.
```

Which is the better model?