A little bit of analysis

Read in the data

Due to the inconsistent column naming covention, we manually convert the cx column to cX in an effort to reduce obfuscation.

```
# Read in the data
data <- read.csv('Greatest_Aussie_Groceries_sales_data.csv', header=TRUE, sep=",")
# Change column name of cx to match style of capital X and Y
colnames(data)[colnames(data)=="cx"] <- "cX"
```

Mutate Data

We mutate the data to include columns for deal_feat (an indictator for both deal and features), revenue, and profit.

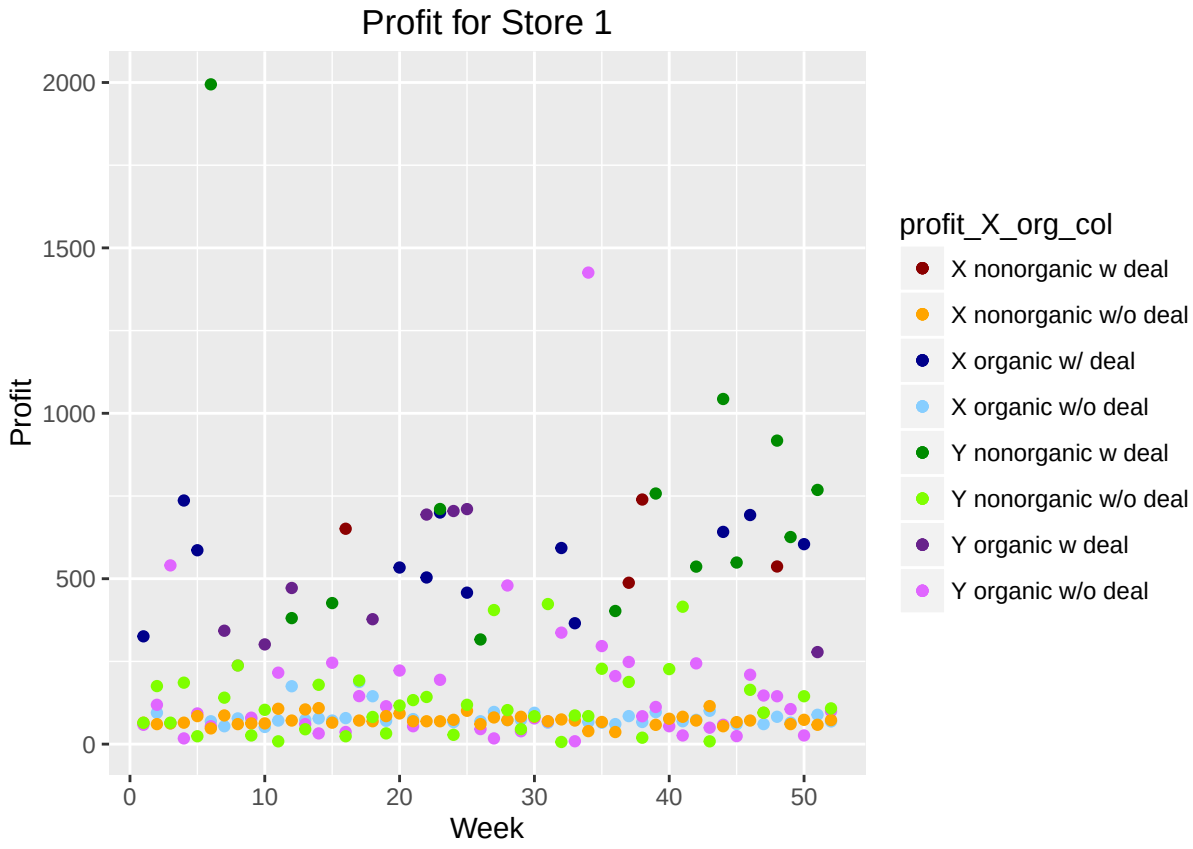
```
# Append a deal_feat column for X and Y
data <- mutate(data, deal_feat_Y = deal_Y*10 + feat_Y, deal_feat_X = deal_X*10 + feat_X)
# Append a revenue column for X and Y
data <- mutate(data, rev_X = oz_X * pX, rev_Y = oz_Y * pY)
# Append a profit column for X and Y
data <- mutate(data, profit_X = rev_X - cX * pX, profit_Y = rev_Y - cY * pY)
```

Summary Plots

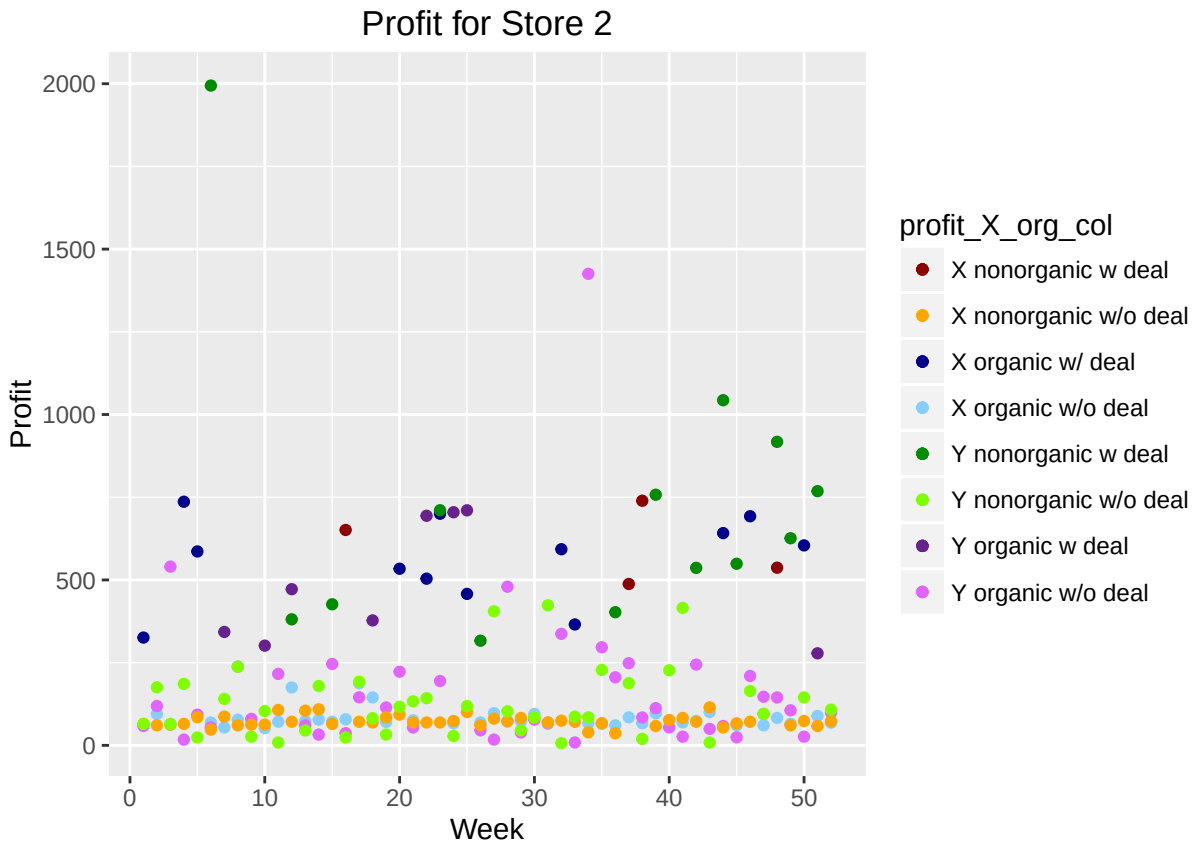
Lets start by just plotting the profit over the 52 weeks for each store.

```
plotStore <- function(STORE) {
# The palette with black:
cbbPalette <- c("red4", "orange1", "blue4", "skyblue1", "green4", "chartreuse1", "darkorchid4", "mediume
# Pull data into temp results dataframe
results <- data.frame(WEEK=c(1:52))
results[,c("profit_X_org","profit_Y_org")] <- data %>% filter(.,STORE==STORE, class=="organic") %>% sele
results[,c("profit_X_non","profit_Y_non")] <- data %>% filter(.,STORE==STORE, class=="nonorganic") %>% s
results[,c("profit_X_org_col","profit_Y_org_col")] <- data %>% filter(.,STORE==STORE, class=="organic")
results[,c("profit_X_non_col","profit_Y_non_col")] <- data %>% filter(.,STORE==STORE, class=="nonorganic
# Assign legend name to categorical data
results$profit_X_org_col[results$profit_X_org_col == 0] <- "X organic w/o deal"
results$profit_X_org_col[results$profit_X_org_col == 1] <- "X organic w/ deal"
results$profit_Y_org_col[results$profit_Y_org_col == 0] <- "Y organic w/o deal"
results$profit_Y_org_col[results$profit_Y_org_col == 1] <- "Y organic w deal"
results$profit_X_non_col[results$profit_X_non_col == 0] <- "X nonorganic w/o deal"
results$profit_X_non_col[results$profit_X_non_col == 1] <- "X nonorganic w deal"
results$profit_Y_non_col[results$profit_Y_non_col == 0] <- "Y nonorganic w/o deal"
results$profit_Y_non_col[results$profit_Y_non_col == 1] <- "Y nonorganic w deal"
# Plot results
ggplot(results, aes(x=WEEK)) +
  geom_point(aes(y=profit_X_org, colour=profit_X_org_col)) +
  geom_point(aes(y=profit_Y_org, colour=profit_Y_org_col)) +
```

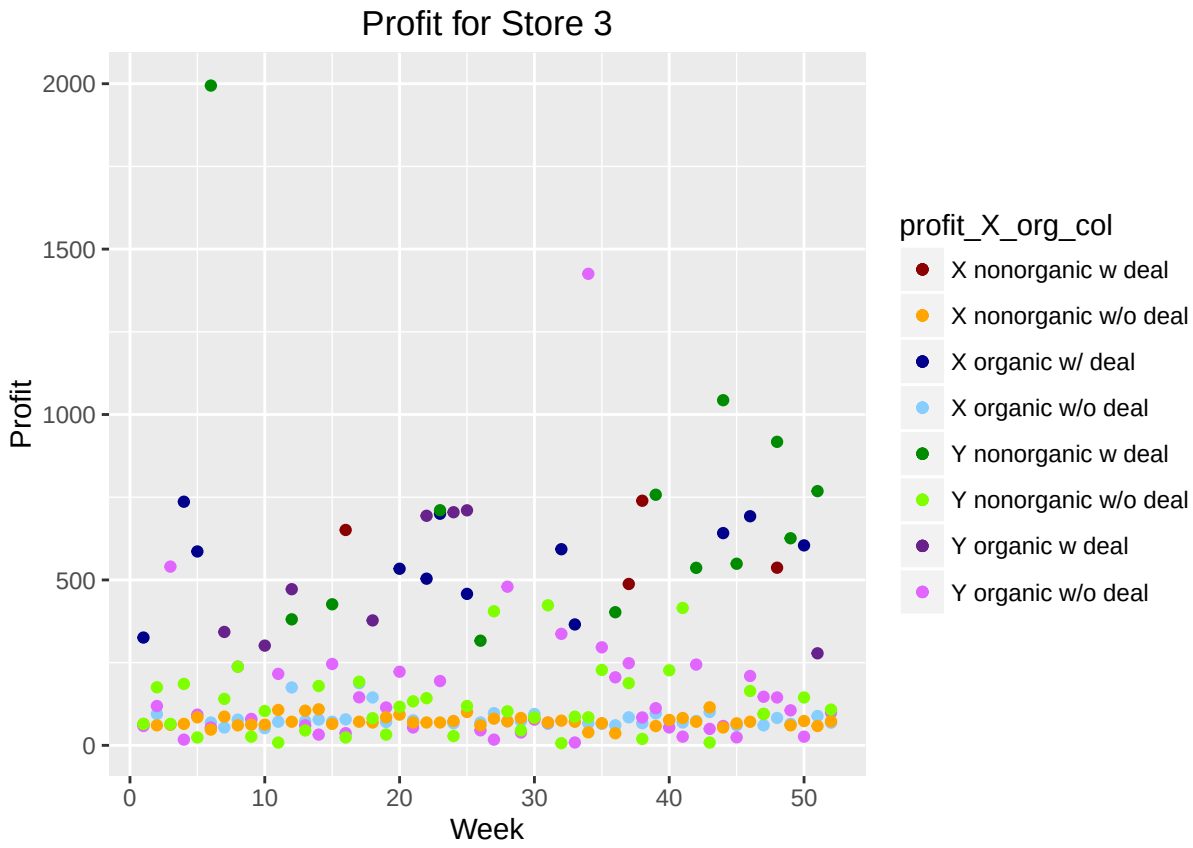
```
  geom_point(aes(y=profit_X_non, colour=profit_X_non_col)) +
  geom_point(aes(y=profit_Y_non, colour=profit_Y_non_col)) +
  scale_colour_manual(values=cbbPalette) +
  labs(x = "Week", y = "Profit", title = paste("Profit for Store",STORE))
}

# Plot the data
plotStore(1)
```



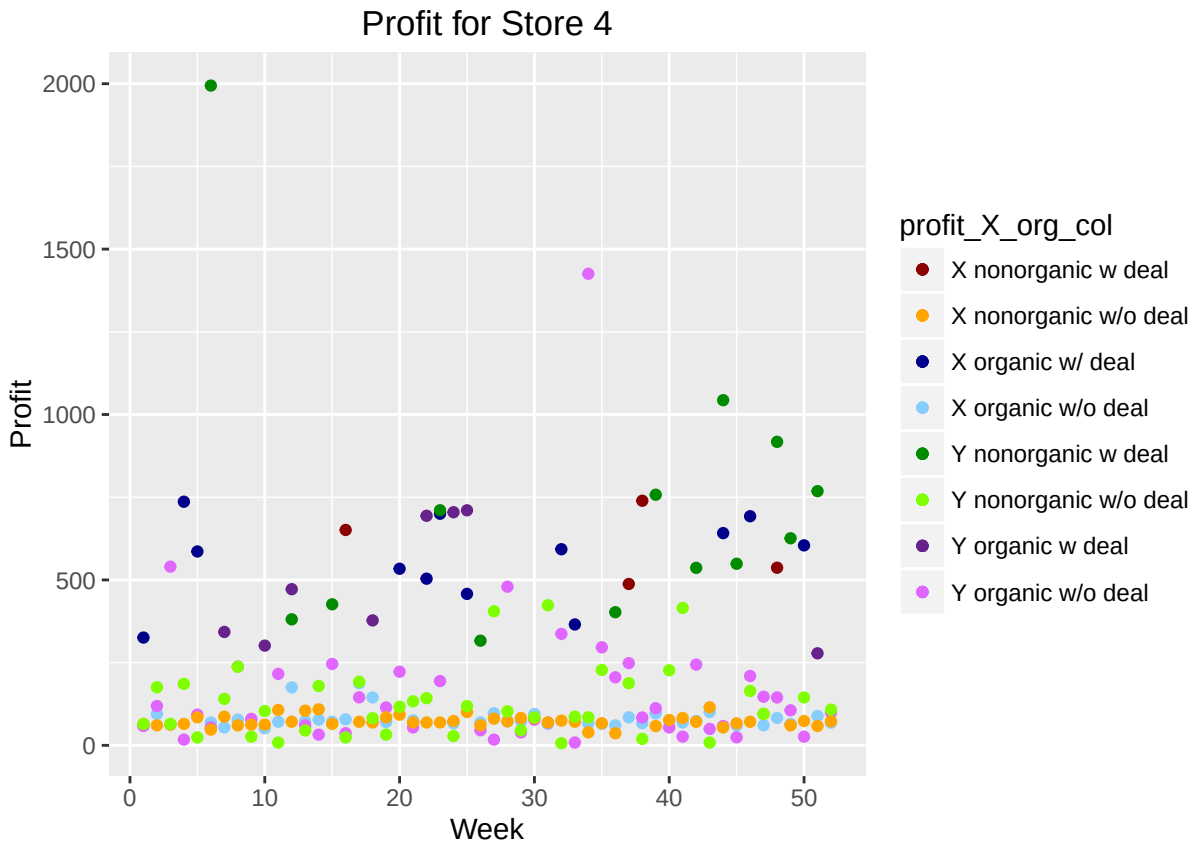

```
plotStore(2)
```


Profit for Store 2
Profit
Week
2000
1500
1000
500
0
0
10
20
30
40
50
profit_X_org_col
X nonorganic w deal
X nonorganic w/o deal
X organic w/ deal
X organic w/o deal
Y nonorganic w deal
Y nonorganic w/o deal
Y organic w deal
Y organic w/o deal

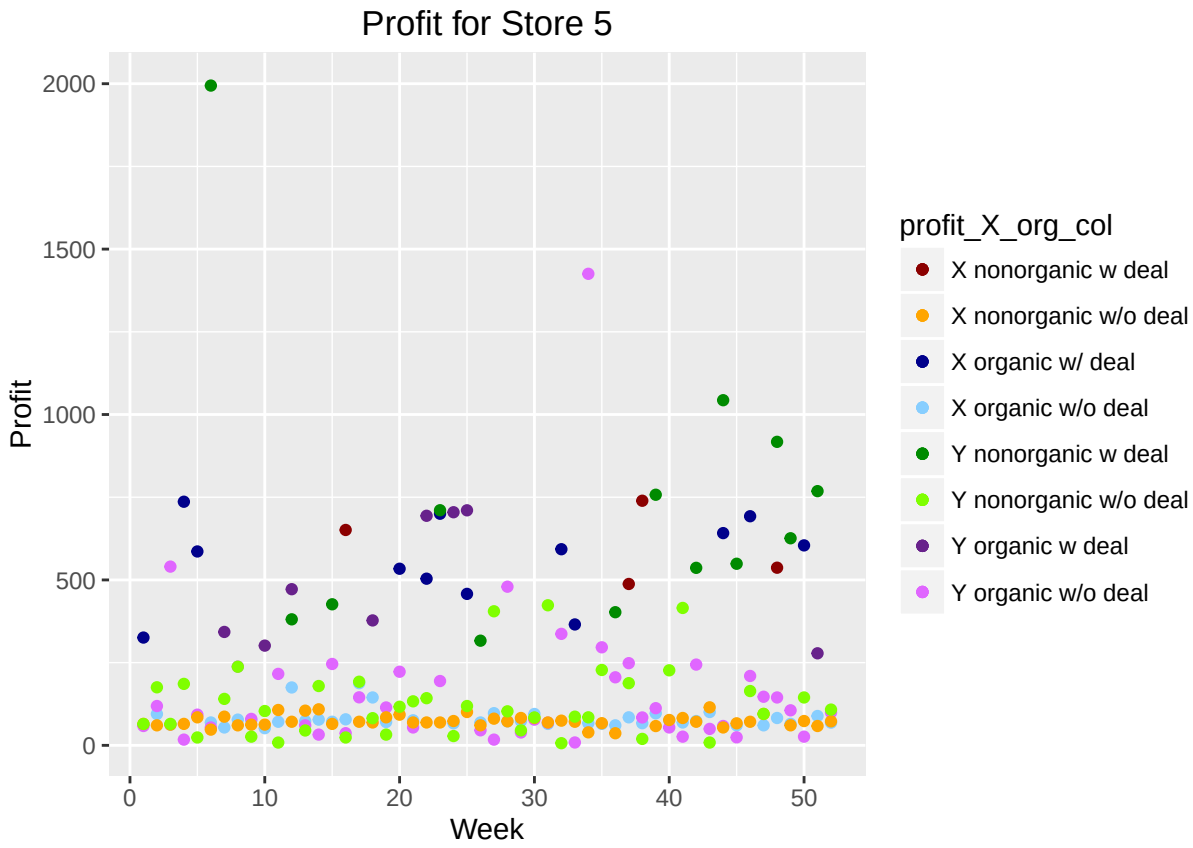
```
plotStore(3)
```


Profit for Store 3
Profit
Week
0
500
1000
1500
2000
0
10
20
30
40
50
profit_X_org_col
X nonorganic w deal
X nonorganic w/o deal
X organic w/ deal
X organic w/o deal
Y nonorganic w deal
Y nonorganic w/o deal
Y organic w deal
Y organic w/o deal

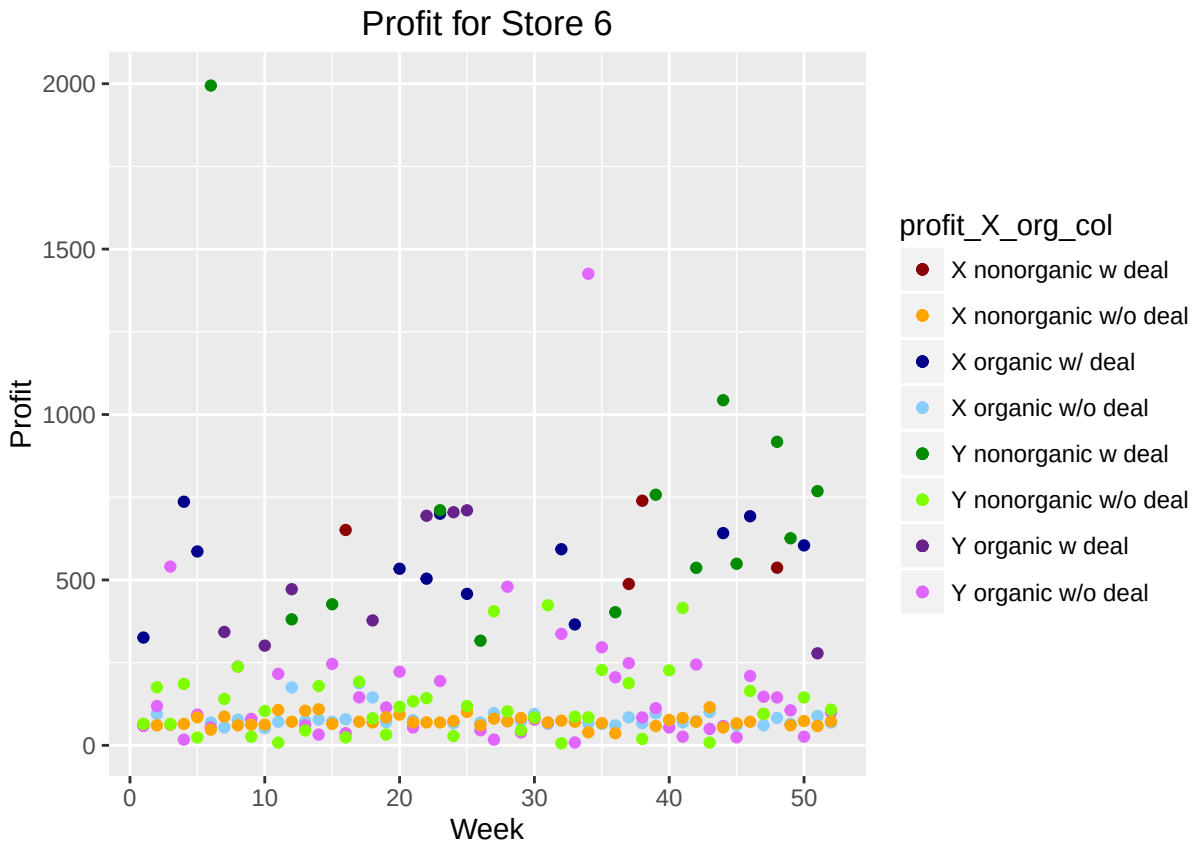
```
plotStore(4)
```


Profit for Store 4
Profit
Week
2000
1500
1000
500
0
0
10
20
30
40
50
profit_X_org_col
X nonorganic w deal
X nonorganic w/o deal
X organic w/ deal
X organic w/o deal
Y nonorganic w deal
Y nonorganic w/o deal
Y organic w deal
Y organic w/o deal

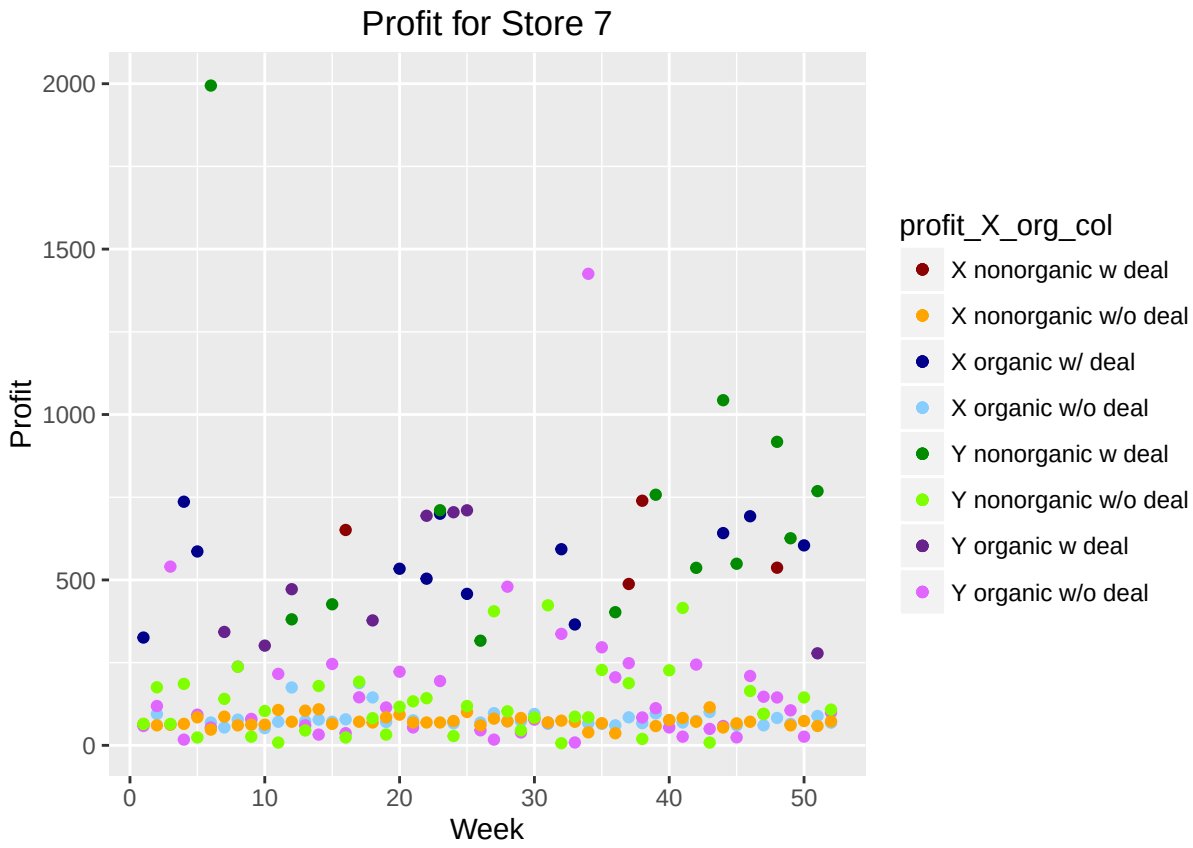
```
plotStore(5)
```


Profit for Store 5
Profit
Week
2000
1500
1000
500
0
0
10
20
30
40
50
profit_X_org_col
X nonorganic w deal
X nonorganic w/o deal
X organic w/ deal
X organic w/o deal
Y nonorganic w deal
Y nonorganic w/o deal
Y organic w deal
Y organic w/o deal

```
plotStore(6)
```


Profit for Store 6
Profit
Week
2000
1500
1000
500
0
0
10
20
30
40
50
profit_X_org_col
X nonorganic w deal
X nonorganic w/o deal
X organic w/ deal
X organic w/o deal
Y nonorganic w deal
Y nonorganic w/o deal
Y organic w deal
Y organic w/o deal

```
plotStore(7)
```

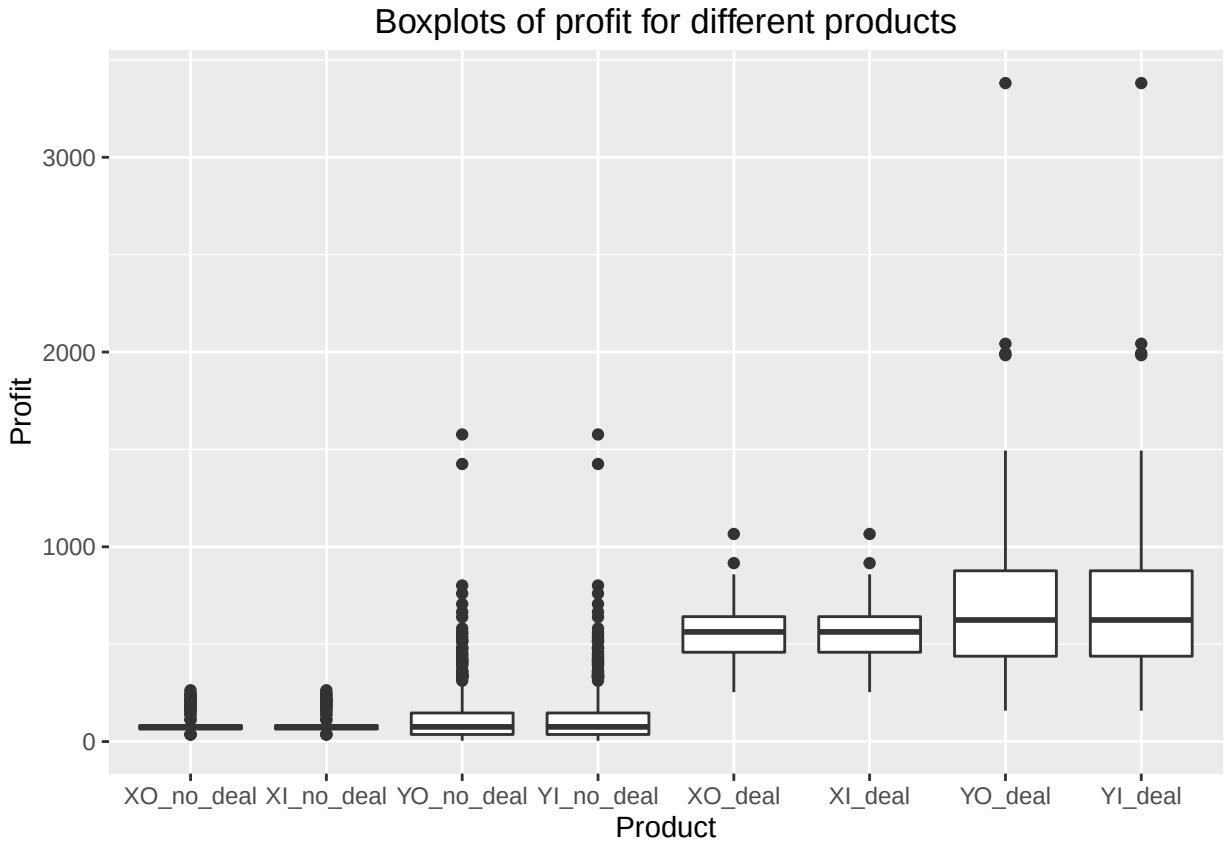


Box plots of profit for the different products

This is just a simple box plot analysis

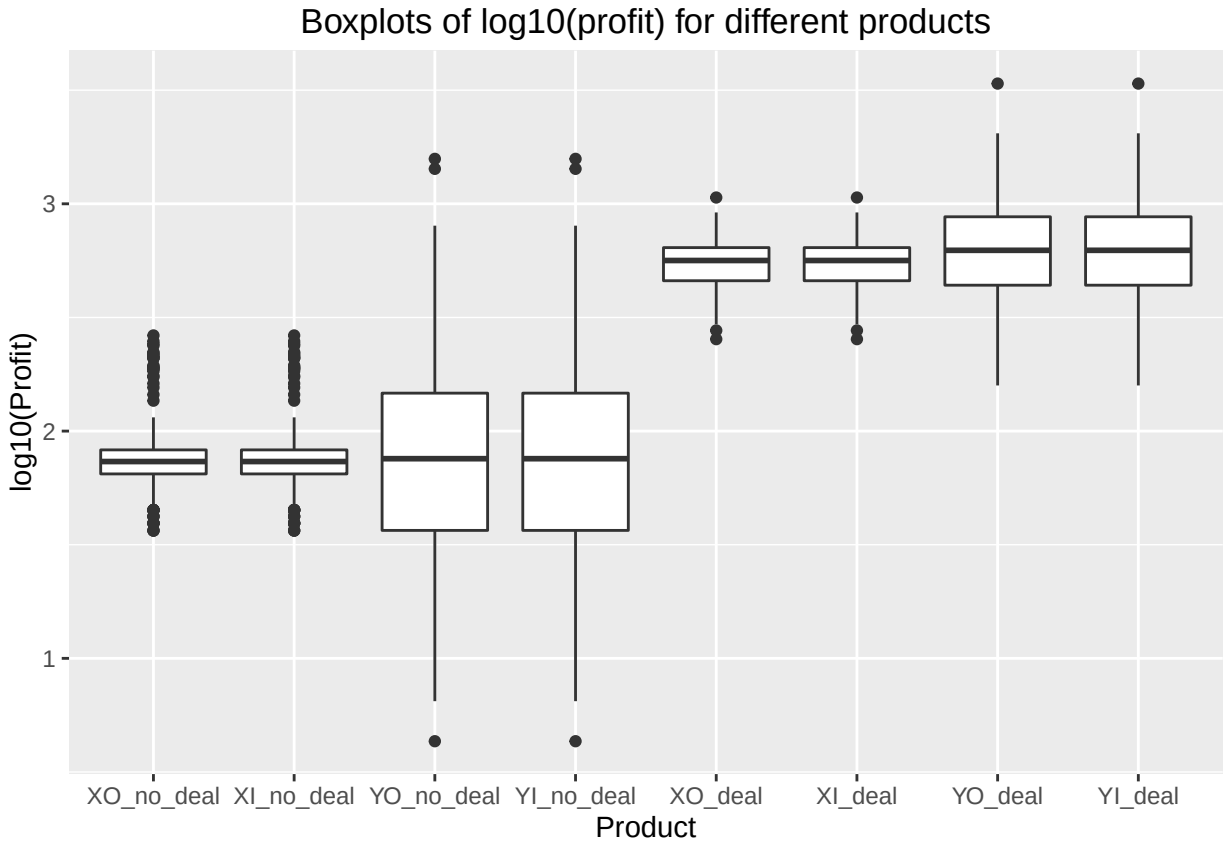
```
# Function to retreve data from dataframe and composite into factor
retrieveData <- function(data, deal, class, xy) {
  names <- c("STORE", "PROFIT")
  if (xy == "x")
    temp <- data %>% filter(deal_X==deal, class==class) %>% select(STORE, profit_X)
   else
    temp <- data %>% filter(deal_Y==deal, class==class) %>% select(STORE, profit_Y)
  colnames(temp) <- names
  name <- if (xy == "x") "X" else "Y"
  name <- if (class == "organic") paste(name,"O",sep="") else paste(name,"I",sep="")
  name <- if (deal == 1) paste(name,"deal",sep="_") else paste(name,"no_deal",sep="_")
  return(data.frame(type=rep(name,nrow(temp)),temp))
}

boxplot_data <- retrieveData(data, deal=0, class="organic", xy="x")
boxplot_data <- rbind(boxplot_data,retrieveData(data, deal=0, class="nonorganic", xy="x"))
boxplot_data <- rbind(boxplot_data,retrieveData(data, deal=0, class="organic", xy="y"))
boxplot_data <- rbind(boxplot_data,retrieveData(data, deal=0, class="nonorganic", xy="y"))
boxplot_data <- rbind(boxplot_data,retrieveData(data, deal=1, class="organic", xy="x"))
boxplot_data <- rbind(boxplot_data,retrieveData(data, deal=1, class="nonorganic", xy="x"))
boxplot_data <- rbind(boxplot_data,retrieveData(data, deal=1, class="organic", xy="y"))
```

```
boxplot_data <- rbind(boxplot_data,retrieveData(data, deal=1, class="nonorganic", xy="y"))

ggplot(boxplot_data, aes(x=type, y=PROFIT)) + geom_boxplot() + labs(title="Boxplots of profit for diffe
```



Given the massive spread between no deal and deal data, let's take a log10() scale transform of the y-axis

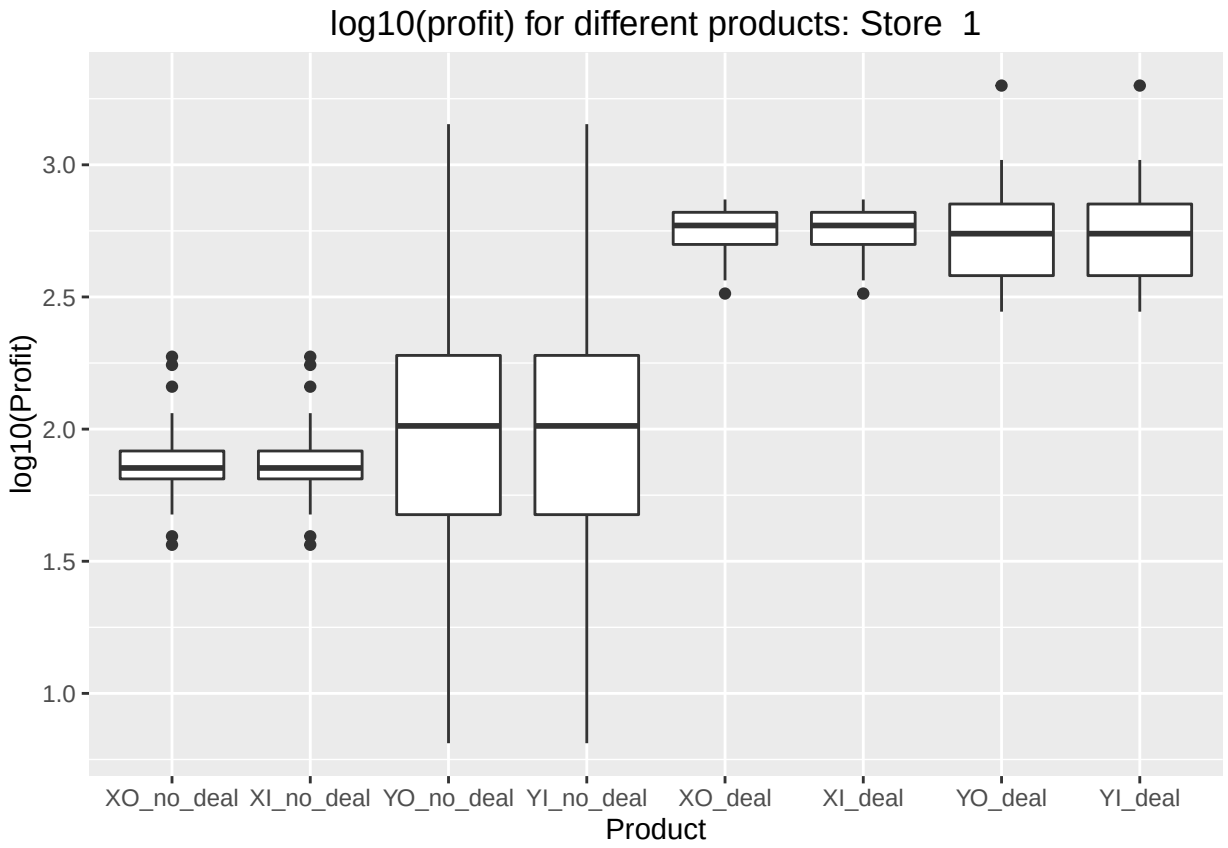
```
boxplot_data <- mutate(boxplot_data, log10PROFIT=log10(PROFIT))
ggplot(boxplot_data, aes(x=type, y=log10PROFIT)) + geom_boxplot() + labs(title="Boxplots of log10(profi
```



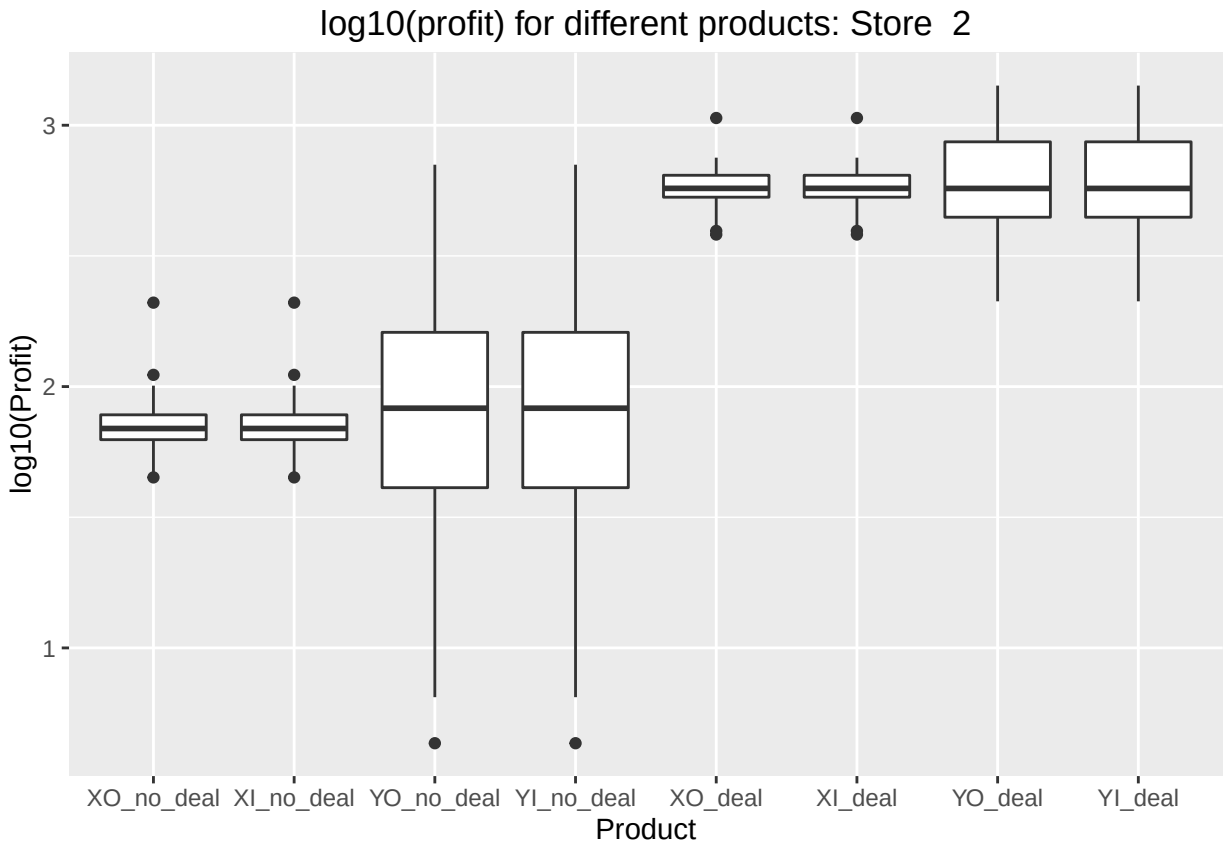
It is clear from these boxplots that the median values for X and Y products are approximately equal. The spread of profit for Y is much larger than X. Profit increases dramatically when a deal is going on.

Just as an additional spam of figures, lets look at the box plots for each product seperated by store (again with a log10 scaling)

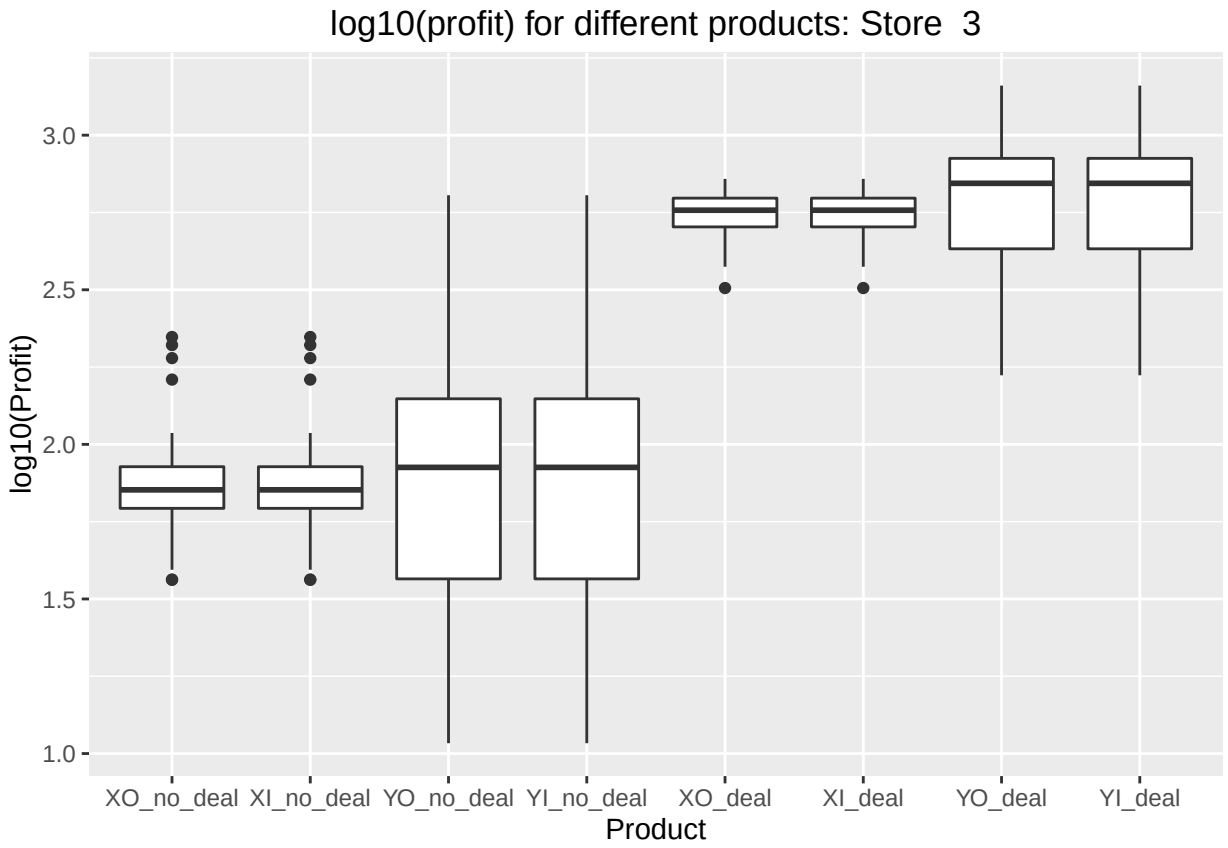
```
boxplot_data %>% filter(STORE==1) %>% ggplot(., aes(x=type, y=log10PROFIT)) + geom_boxplot() + labs(tit
```



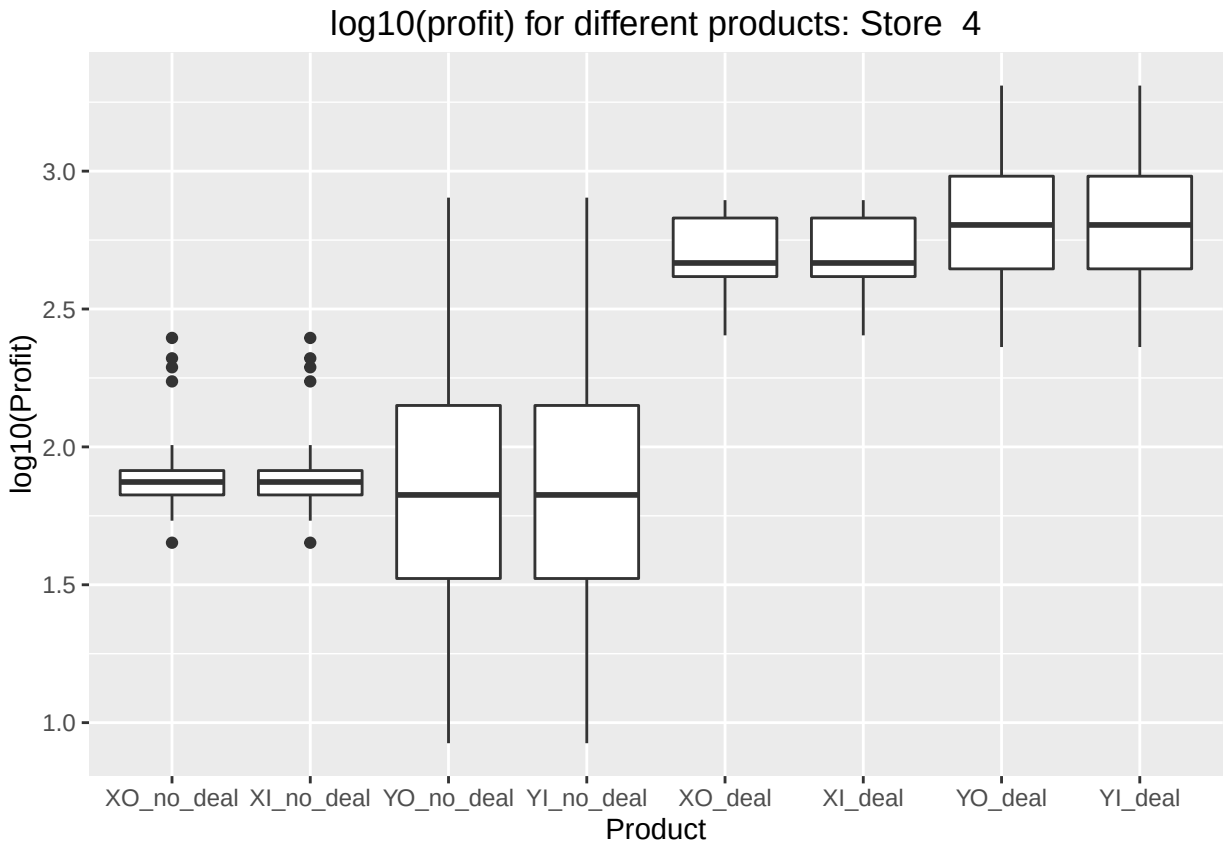

```
boxplot_data %>% filter(STORE==2) %>% ggplot(., aes(x=type, y=log10PROFIT)) + geom_boxplot() + labs(tit
```



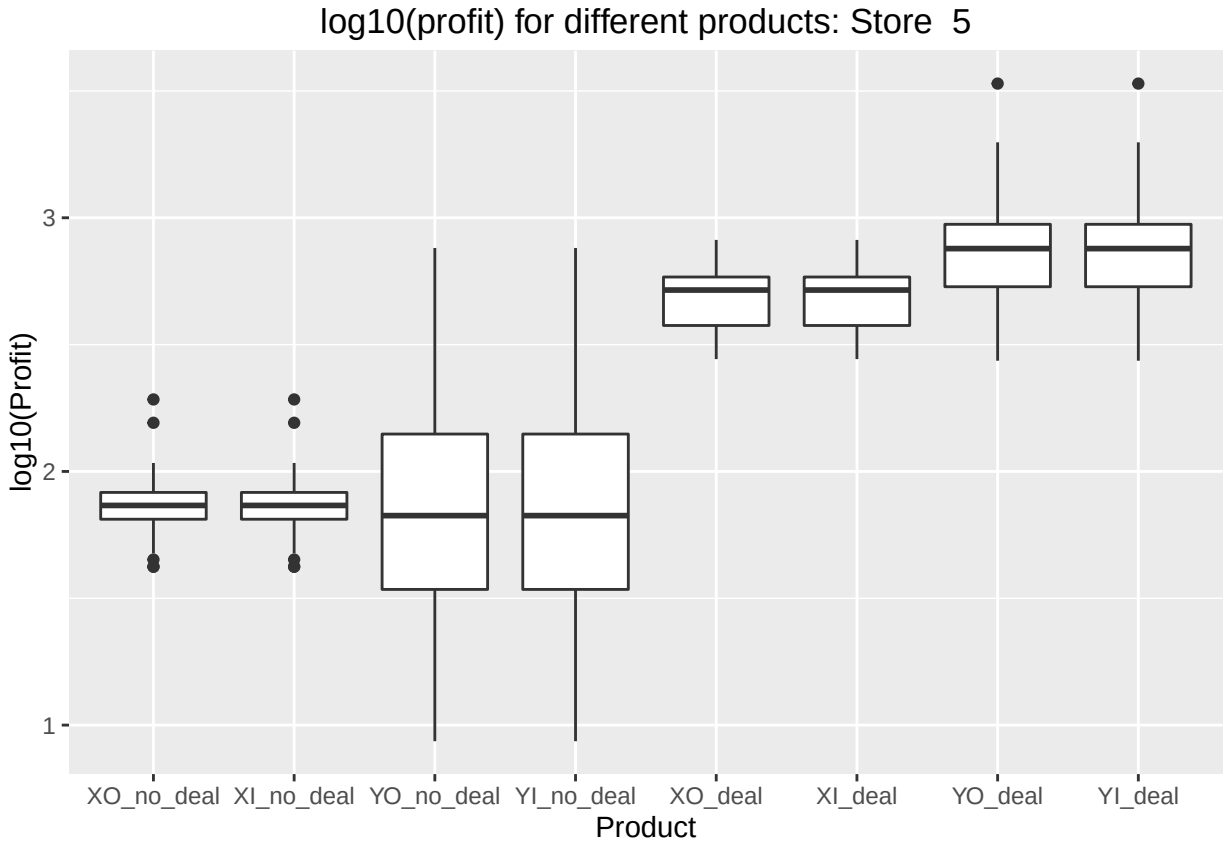

```
boxplot_data %>% filter(STORE==3) %>% ggplot(., aes(x=type, y=log10PROFIT)) + geom_boxplot() + labs(tit
```




```
boxplot_data %>% filter(STORE==4) %>% ggplot(., aes(x=type, y=log10PROFIT)) + geom_boxplot() + labs(tit
```

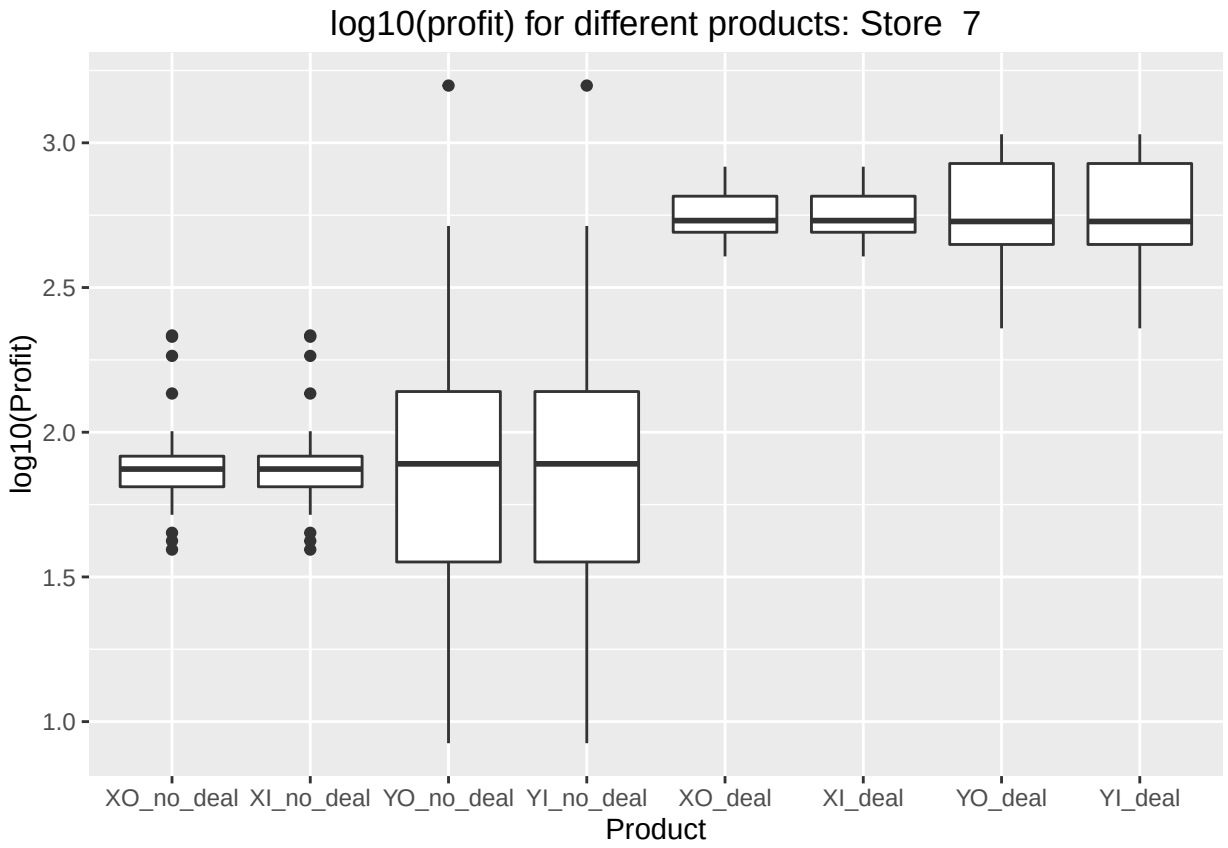



```
boxplot_data %>% filter(STORE==5) %>% ggplot(., aes(x=type, y=log10PROFIT)) + geom_boxplot() + labs(tit
```




```
boxplot_data %>% filter(STORE==6) %>% ggplot(., aes(x=type, y=log10PROFIT)) + geom_boxplot() + labs(tit
```


```
boxplot_data %>% filter(STORE==7) %>% ggplot(., aes(x=type, y=log10PROFIT)) + geom_boxplot() + labs(tit
```


log10(profit) for different products: Store 7
log10(Profit)
3.0
2.5
2.0
1.5
1.0
XO_no_deal
XI_no_deal
YO_no_deal
YI_no_deal
XO_deal
XI_deal
YO_deal
YI_deal
Product

Regression