Final Project

Brock Bradfield

06/30/2022

Packages, Directory, and Data

```
library(tidyr)
library(dplyr)
library(readr)
library(ggplot2)

setwd("~/DA/R")


invoice_lines <- read_csv('invoice_lines_dirty.csv')
```

Data Summary

```
summary(invoice_lines)
```

```
##  InvoiceLineID      InvoiceID      StockItemID    Description
##  Min.   :     1   Min.   :    1   Min.   :  1.0   Length:228265
##  1st Qu.: 57067   1st Qu.:17572   1st Qu.: 54.0   Class :character
##  Median :114133   Median :35152   Median :111.0   Mode  :character
##  Mean   :114133   Mean   :35179   Mean   :110.2
##  3rd Qu.:171199   3rd Qu.:52765   3rd Qu.:165.0
##  Max.   :228265   Max.   :70510   Max.   :227.0
##                                   NA's   :141
##  PackageTypeID       Quantity        UnitPrice          TaxRate
##  Min.   : 1.000   Min.   :  1.00   Min.   :   0.00   Min.   :-15.00
##  1st Qu.: 7.000   1st Qu.:  5.00   1st Qu.:  13.00   1st Qu.: 15.00
##  Median : 7.000   Median : 10.00   Median :  18.00   Median : 15.00
##  Mean   : 7.074   Mean   : 39.21   Mean   :  45.59   Mean   : 14.98
##  3rd Qu.: 7.000   3rd Qu.: 60.00   3rd Qu.:  32.00   3rd Qu.: 15.00
##  Max.   :10.000   Max.   :360.00   Max.   :1899.00   Max.   : 15.00
##  NA's   :24                        NA's   :79
##    TaxAmount        LineProfit     ExtendedPrice       LastEditedBy
##  Min.   :   0.0   Min.   :-645.0   Min.   :     0.0   Min.   : 2.0
##  1st Qu.:  14.4   1st Qu.:  51.0   1st Qu.:   110.4   1st Qu.: 6.0
##  Median :  34.5   Median : 120.0   Median :   264.5   Median :11.0
##  Mean   : 112.9   Mean   : 375.6   Mean   :   868.4   Mean   :10.8
##  3rd Qu.: 129.6   3rd Qu.: 390.0   3rd Qu.:   993.6   3rd Qu.:16.0
##  Max.   :2848.5   Max.   :9200.0   Max.   :218385.0   Max.   :20.0
##                                                      NA's   :177849
##  LastEditedWhen
##  Length:228265
```

```
##  Class :character
##  Mode  :character
##
##
##
##
```

Data Cleaning

```
# Shows a count of the NAs for each column
na_count <- colSums(is.na(invoice_lines))
print(na_count)
```

```
##  InvoiceLineID      InvoiceID    StockItemID    Description  PackageTypeID
##              0              0            141            140             24
##       Quantity      UnitPrice        TaxRate      TaxAmount     LineProfit
##              0             79              0              0              0
##  ExtendedPrice   LastEditedBy LastEditedWhen
##              0         177849              0
```

```
# Shows the total number of rows in the dataset
total_rows <- nrow(invoice_lines)
print(total_rows)
```

```
## [1] 228265
```

```
# Drops the LastEditedBy column (shows new na count)
invoice_lines <- subset(invoice_lines, select = -LastEditedBy)
new_na_count <- colSums(is.na(invoice_lines))
print(new_na_count)
```

```
##  InvoiceLineID      InvoiceID    StockItemID    Description  PackageTypeID
##              0              0            141            140             24
##       Quantity      UnitPrice        TaxRate      TaxAmount     LineProfit
##              0             79              0              0              0
##  ExtendedPrice LastEditedWhen
##              0              0
```

```
# Imputes missing unit prices based on the description-to-price associations
invoice_lines <- invoice_lines %>%
  mutate(UnitPrice = ifelse(is.na(UnitPrice),
                            match(Description, invoice_lines$Description) %>%
                              `[`(., UnitPrice),
                            UnitPrice))

new2_na_count <- colSums(is.na(invoice_lines))
print(new2_na_count)
```

```
##  InvoiceLineID      InvoiceID    StockItemID    Description  PackageTypeID
##              0              0            141            140             24
##       Quantity      UnitPrice        TaxRate      TaxAmount     LineProfit
##              0              0              0              0              0
##  ExtendedPrice LastEditedWhen
##              0              0
```

```
# Drops the na rows in the 'StockItemID' column
invoice_lines <- invoice_lines %>%
```

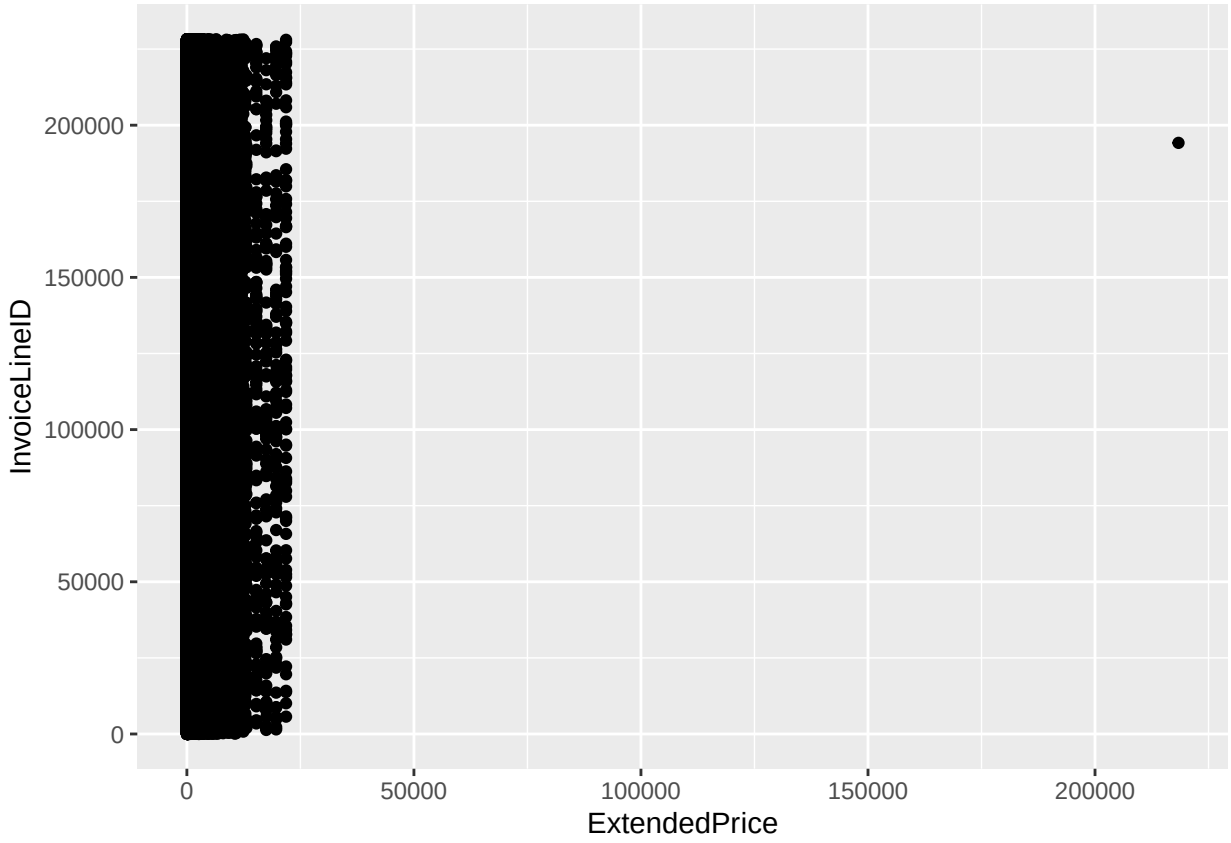
```
  drop_na(StockItemID)

# Drops the na rows in the 'PackageTypeID' column
invoice_lines <- invoice_lines %>%
  drop_na(PackageTypeID)

# Shows the new count of na's in each column
new3_na_count <- colSums(is.na(invoice_lines))
print(new3_na_count)
```

```
##  InvoiceLineID      InvoiceID    StockItemID    Description  PackageTypeID
##              0              0              0              0              0
##       Quantity      UnitPrice        TaxRate      TaxAmount     LineProfit
##              0              0              0              0              0
##  ExtendedPrice LastEditedWhen
##              0              0
```

Stakeholder Question 1

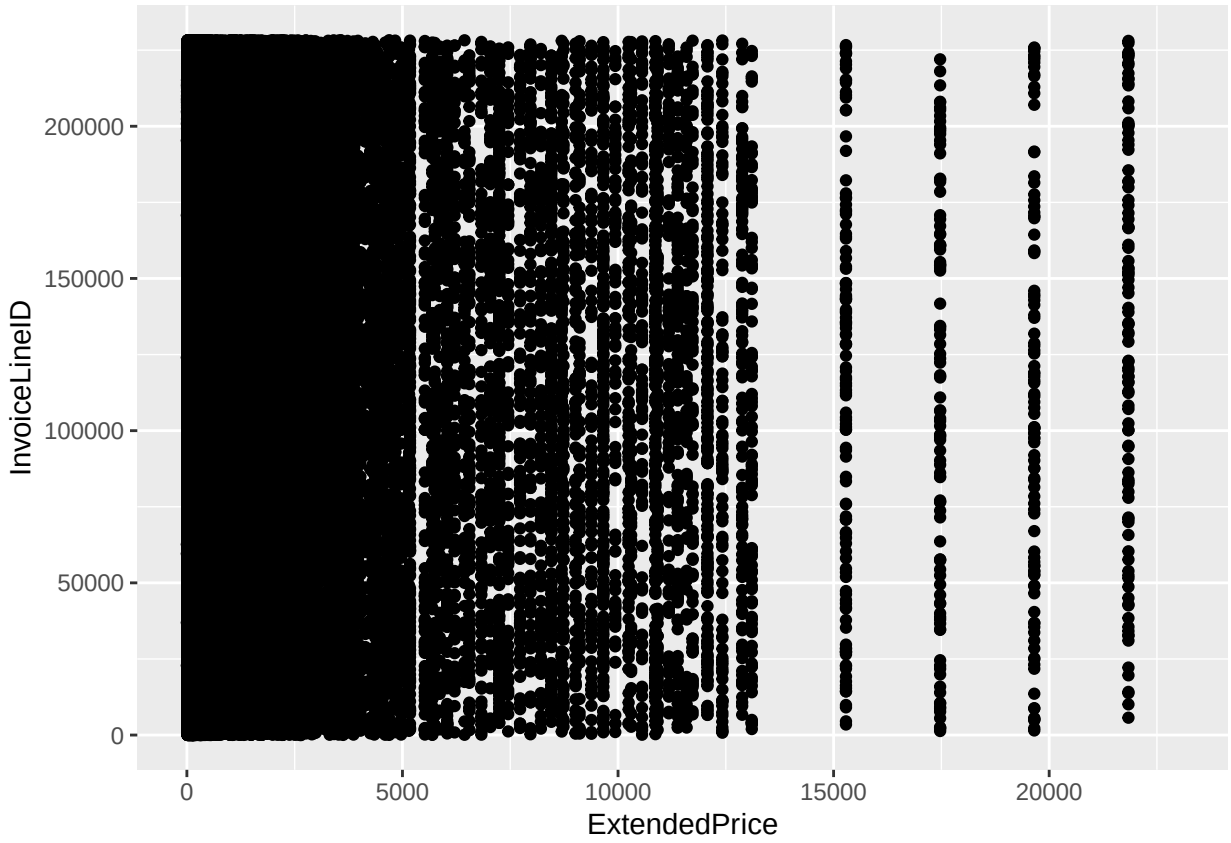
What is the spending pattern on invoices?

```
# Scatter plot (ExtendedPrice)
ggplot(data=invoice_lines) +
  geom_point(mapping=aes(x=ExtendedPrice, y=InvoiceLineID))
```

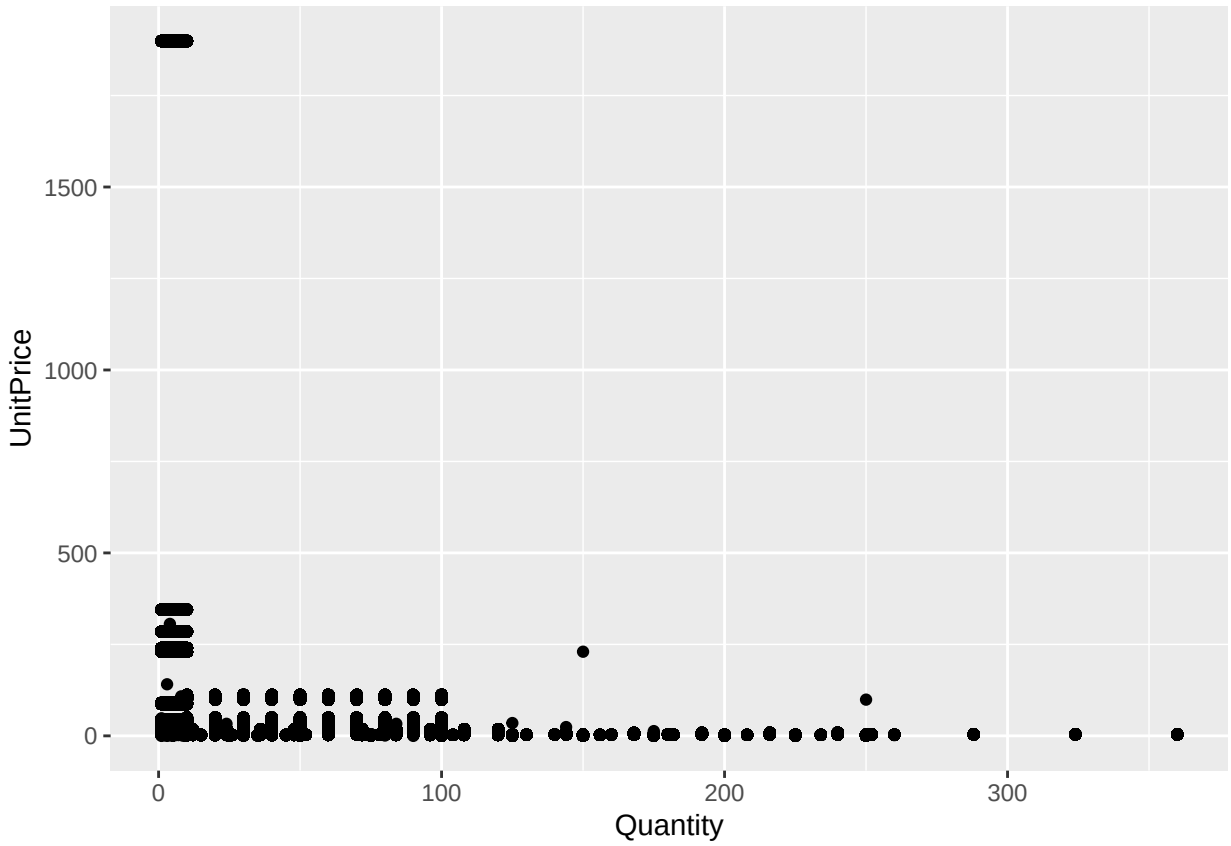


```
ggplot(data=invoice_lines) +
  geom_point(mapping=aes(x=ExtendedPrice, y=InvoiceLineID)) +
  scale_x_continuous(limits=c(0, 23000))
```

```
## Warning: Removed 1 row containing missing values or values outside the scale range
## (`geom_point()`).
```




```
# Scatter plot (Quantity & UnitPrice)
ggplot(data =invoice_lines) +
  geom_point(mapping=aes(x=Quantity, y=UnitPrice))
```



\textit{When looking at the first graph it's kind of hard to notice any significance. I only included it to show the one oulier. Invoice line id somehwere around 190,000 had an extended of price of close to $225,000 when none other points went outside of 25,000, which I thought was interesting because that was a really big purchase. The second graph has a little more significance. We can see more of what's going on with the spending pattern of the invoices. Unsurspringly we can see that is it most common to see purchases in the 0-3,000 range. It is less common to see purchases in the 5,000 - 13,000 range, but still farely common, and least common to see purchases in the 15,000 - 25,000 range. The third graph shows us that the higher the qantity the lower the unit price was, or the lower the quanity, the higher the unit price was. This makes sense to me as most people would probably purchase less quantity of the more expensive things.}

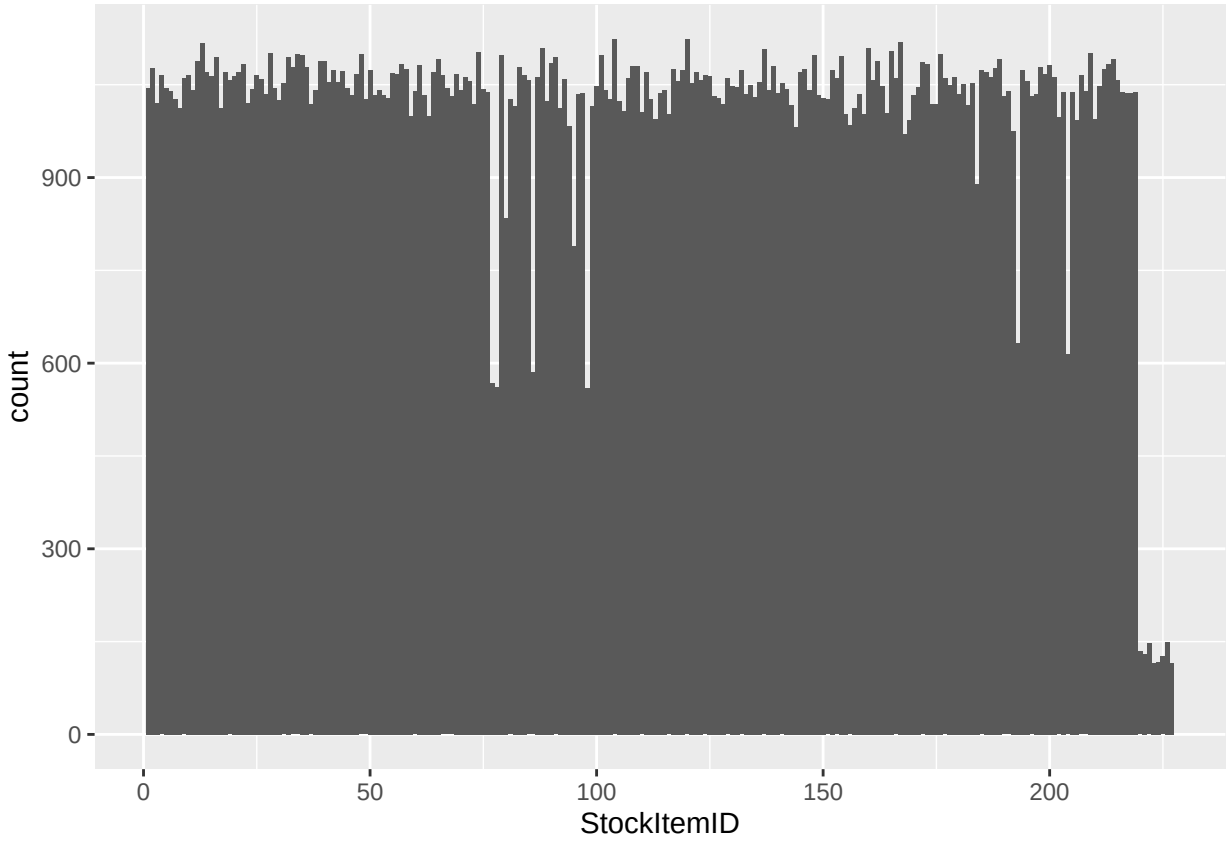
Stakeholder Question 2

What are the top 5 most popular stock items?

```
# Counts occurrences of each StockItemID using the table function
stockitem_counts <- table(invoice_lines$StockItemID)

# Converts the table to a dataframe
stockitem_counts_df <- as.data.frame(stockitem_counts)

# Sorts frequency in descending order
sorted_stockitem_counts <- stockitem_counts_df[order(-stockitem_counts_df$Freq), ]

# Gets the top 5 most popular StockItemID's
top_5 <- head(sorted_stockitem_counts, 5)
print(top_5)
```

```
##     Var1 Freq
## 104  104 1123
## 120  120 1123
## 167  167 1119
## 13    13 1117
## 88    88 1109
```

```
# Bar graph (StockItemID column)
ggplot(data=invoice_lines) +
  geom_bar(mapping=aes(x=StockItemID))
```



Stock item id 104, 'Alien officer hoodie (Black) 3XL' were the most popular, purhcased a total of 1,123 times. Stock item id 120, 'Dinosaur battery-powered slippers (Green) L' were the second most popular, purchased a total of 1,123 times. Stock item id 167, '10 mm Anti static bubble wrap (Blue) 50m' were the third most popular, purchased a total of 1,119 times. Stock item id 13, 'USB food flash drive - shrimp cocktail' were the fourth most popular, purchased a total of 1,117 times. And stock item id 88, '"The Gu" red shirt XML tag t-shirt (White) 7XL' were the fifth most popular, purchased a total of 1,109 times. However when looking at the graph I think it's worth mentioning that although yes these are the top 5 most popular stock items, the vast majority of the stock items were sitting close to that range of total number of times purchased.

What are the 5 least popular stock items?

```
# Sorts frequency in ascending order
ascending_stockitem_counts <- stockitem_counts_df[order(stockitem_counts_df$Freq), ]

# Gets the top 5 least popular StockItemID's
top_5_least_popular <- head(ascending_stockitem_counts, 5)
print(top_5_least_popular)
```

```
##     Var1 Freq
## 223  223  115
## 227  227  116
## 224  224  117
## 225  225  126
## 221  221  130
```

According to the results of the code above, stock item id 223, 'Chocolate echidnas 250g' were the least popular, purchased a total of 115 times. Stock item id 227, 'White chocolate moon rocks 250g' were the second least popular, purchased a total of 116 times. Stock item id 224, 'Chocolate frogs 250g' were the third least popular, purchased a total of 117 times. Stock item id, 225, 'Chocolate sharks 250g' were the fourth least popular, purchased a total of 126 times. And stock item id 221, 'Novelty chilli chocolates 500g' were the fifth least popular, purchased a total of 130 times.

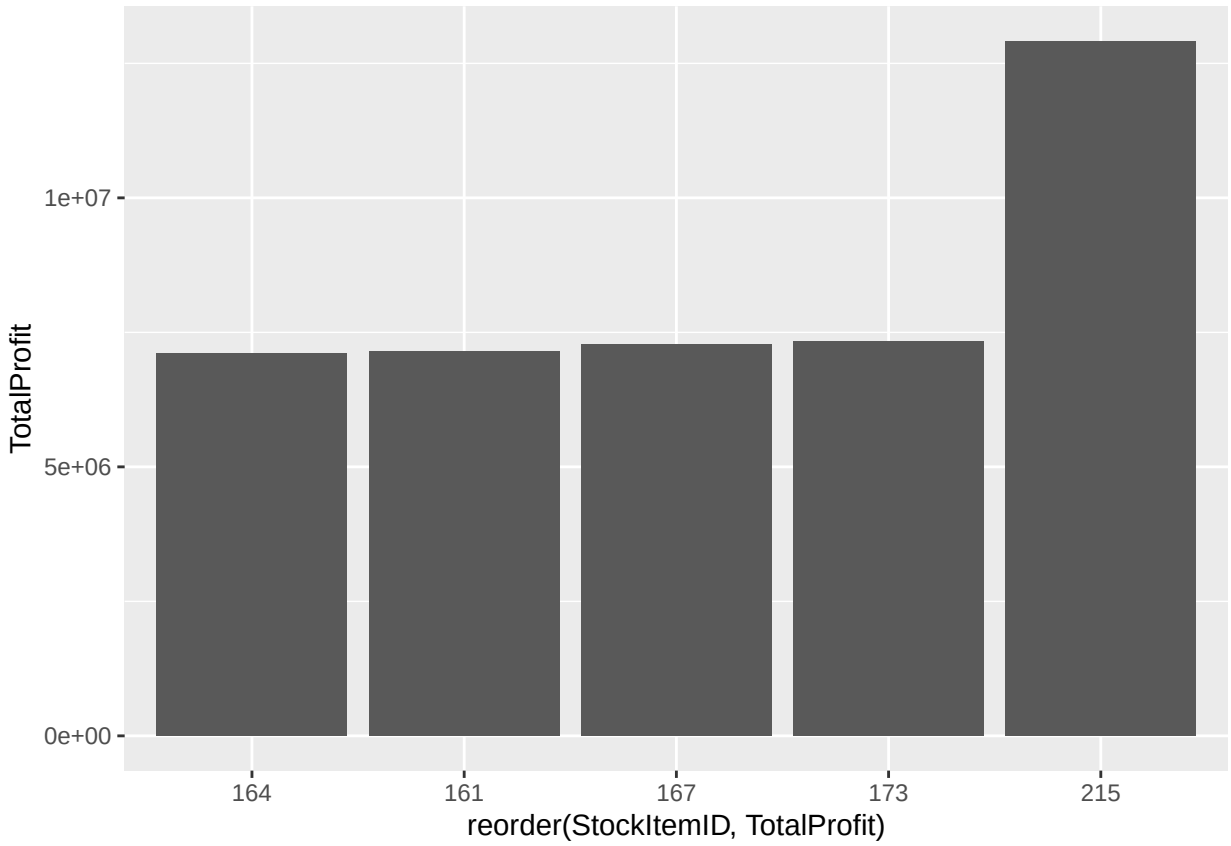
Stakeholder Question 3

Which 5 stock items bring in the most gross profit?

```r
# Groups by StockItemID and sums the ExtendedPrice
profit_by_item <- invoice_lines %>%
  group_by(StockItemID) %>%
  summarize(TotalProfit = sum(ExtendedPrice))

# Sorts the results by total profit in descending order
sorted_profit <- profit_by_item %>%
  arrange(desc(TotalProfit))

# Selects the top 5 stock items
top_5_profitable_items <- head(sorted_profit, 5)

# Prints the top 5 profitable items
print(top_5_profitable_items)
```

```
## # A tibble: 5 x 2
##   StockItemID TotalProfit
##         <dbl>       <dbl>
## 1         215   12915289.
## 2         173    7324695
## 3         167    7268184
## 4         161    7143984
## 5         164    7118776
```

```r
# Bar graph (top_5_profitable_items)
ggplot(data = top_5_profitable_items, aes(x = reorder(StockItemID, TotalProfit), y = TotalProfit)) +
  geom_bar(stat = "identity")
```



\textit{Stock item id 215, 'Air cushion machine (Blue)' had the highest total profit at $12,915,289. Stock item id 173, '32 mm Anti static bubble wrap (Blue) 50m' had the second highest total profit at $7,324,695. Stock item id 167, '10 mm Anti static bubble wrap (Blue) 50m' had the third highest total profit at $7,268,184. Stock item id 161, '20 mm Double sided bubble wrap 50m' had the fourth highest profit at $7,143,984. And stock item id 164, '32 mm Double sided bubble wrap 50m' had the fifth highest profit at $7,118,776.}

Which 5 stock items bring in the least gross profit?

```
#Sorts the results by total profit in ascending order
ascending_sorted_profit <- profit_by_item %>%
  arrange(TotalProfit)

# Selects the top 5 stock items with the least amount of total profit
least_5_profitable_items <- head(ascending_sorted_profit, 5)

# Prints the top 5 stock items with the least amount of total profit
print(least_5_profitable_items)
```

```
## # A tibble: 5 x 2
##   StockItemID TotalProfit
##         <dbl>       <dbl>
## 1         209      65968.
## 2         210      74603.
## 3          17      80760.
## 4          27      81791.
## 5          43      83047.
```

According to the code above we can see that stock item id 209, 'Packing knife with metal insert blade (Yellow)

9mm' had the least amount of total profit. Stock item id 210, 'Packing knife with metal insert blade (Yellow) 18mm' had the second least amount of total profit. Stock item id 17, 'DBA joke mug - mind if I join you? (Black)' had the third least amount of total profit. Stock item id 27, 'DBA joke mug - SELECT caffeine FROM mug (Black)' had the fourth least amount of total profit. And stock item id 43, 'Developer joke mug - understanding recursion requires understanding recursion (Black)' had the fifth least amount of total profit.

Stakeholder Question 4

Which stock items do you recommend focusing on?

\textit{The stock items I would recommend to focus on first on foremost would be Stock items 215, 173, 167, 161, 164 as these were the top 5 stock items that brought in the most total profit. I would also focus on stock items 104, 120, 167, 13, 88 as these were the top 5 most popular stock items. Surprisingly there was only one overlap between these two analysis 167, which might be a really good stock item to focus on.

215: Air cushion machine (Blue) 173: 32 mm Anti static bubble wrap (Blue) 50m 167: 10 mm Anti static bubble wrap (Blue) 50m 161: 20 mm Double sided bubble wrap 50m 164: 32 mm Double sided bubble wrap 50m

104: Alien officer hoodie (Black) 3XL 120: Dinosaur battery-powered slippers (Green) L 167: 10 mm Anti static bubble wrap (Blue) 50m 13: USB food flash drive - shrimp cocktail 88: “The Gu” red shirt XML tag t-shirt (White) 7XL