Midterm

Cody Evans

Importing data

I first downloaded the data but sadly couldn't get a working web link show other could easily replicate or mess with what I've done.

I started off by importing the libraries and data I would be using.

```r
suppressMessages(library("tidyverse"))
suppressMessages(register <- read_csv("register.csv"))
suppressMessages(sales <- read_csv("sales.csv"))
#N <- nrow(register)
```

Original column names for the two data sets Register and Sales

Register

```r
colnames(register)
```

```
##  [1] "purchase"
##  [2] "item"
##  [3] "charge"
##  [4] "price"
##  [5] "is the customer a student/faculty/staff (0) or unaffiliated (1)"
##  [6] "customer id"
##  [7] "receipt"
##  [8] "contact preference"
##  [9] "newsletter"
## [10] "sales"
## [11] "preferred customer discount"
```

Sales

```r
colnames(sales)
```

```
##  [1] "category of inventory goods" "1-2018"
##  [3] "2-2018"                      "3-2018"
##  [5] "4-2018"                      "5-2018"
##  [7] "6-2018"                      "7-2018"
##  [9] "8-2018"                      "9-2018"
## [11] "10-2018"                     "11-2018"
## [13] "12-2018"                     "1-2019"
## [15] "2-2019"                      "3-2019"
## [17] "4-2019"                      "5-2019"
## [19] "6-2019"                       "7-2019"
## [21] "8-2019"                       "9-2019"
## [23] "10-2019"
```

Changing column names for better understanding

Register

cid is the new customer id. pid is the purchase id. And I put unaffiliated to cut down on the size of the name but retain understanding

```
colnames(register)[colnames(register) == "is the customer a student/faculty/staff (0) or unaffiliated (
colnames(register)[colnames(register) == "customer id"] <- "cid"
colnames(register)[colnames(register) == "purchase"] <- "pid"
colnames(register)[colnames(register) == "preferred customer discount"] <- "discount"
colnames(register)[colnames(register) == "contact preference"] <- "contact"

colnames(register)
```

```
##  [1] "pid"          "item"         "charge"       "price"
##  [5] "unaffiliated" "cid"          "receipt"      "contact"
##  [9] "newsletter"   "sales"        "discount"
```

Sales

goods is the type of goods.

```
colnames(sales)[colnames(sales) == "category of inventory goods"] <- "goods"

colnames(sales)
```

```
##  [1] "goods"   "1-2018"  "2-2018"  "3-2018"  "4-2018"  "5-2018"  "6-2018"
##  [8] "7-2018"  "8-2018"  "9-2018"  "10-2018" "11-2018" "12-2018" "1-2019"
## [15] "2-2019"  "3-2019"  "4-2019"  "5-2019"  "6-2019"  "7-2019"  "8-2019"
## [22] "9-2019"  "10-2019"
```

Transforming the dataset tibbles to slightly better tibbles

And setting the data types to factors if they should be factors like if the charge was tax or not

```
purchases <- tibble(pid = register$pid, item = as.factor(register$item), charge = as.factor(register$cha

sales$goods <- as.factor(sales$goods)
```

Example

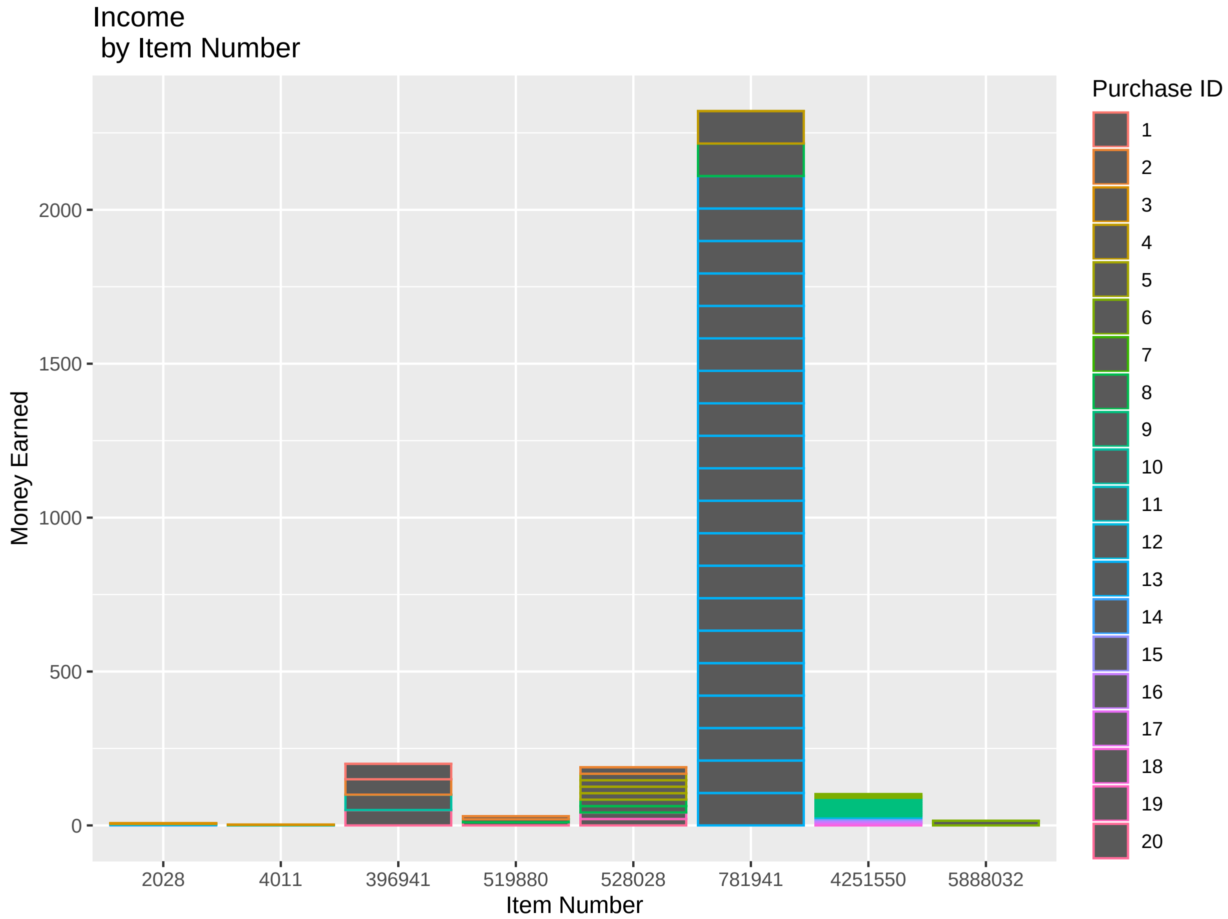
This is an example of something that you could look into

This is showing how much a certain item brought in from the register table

Neither of these tables show tax as that is not money made by the store.

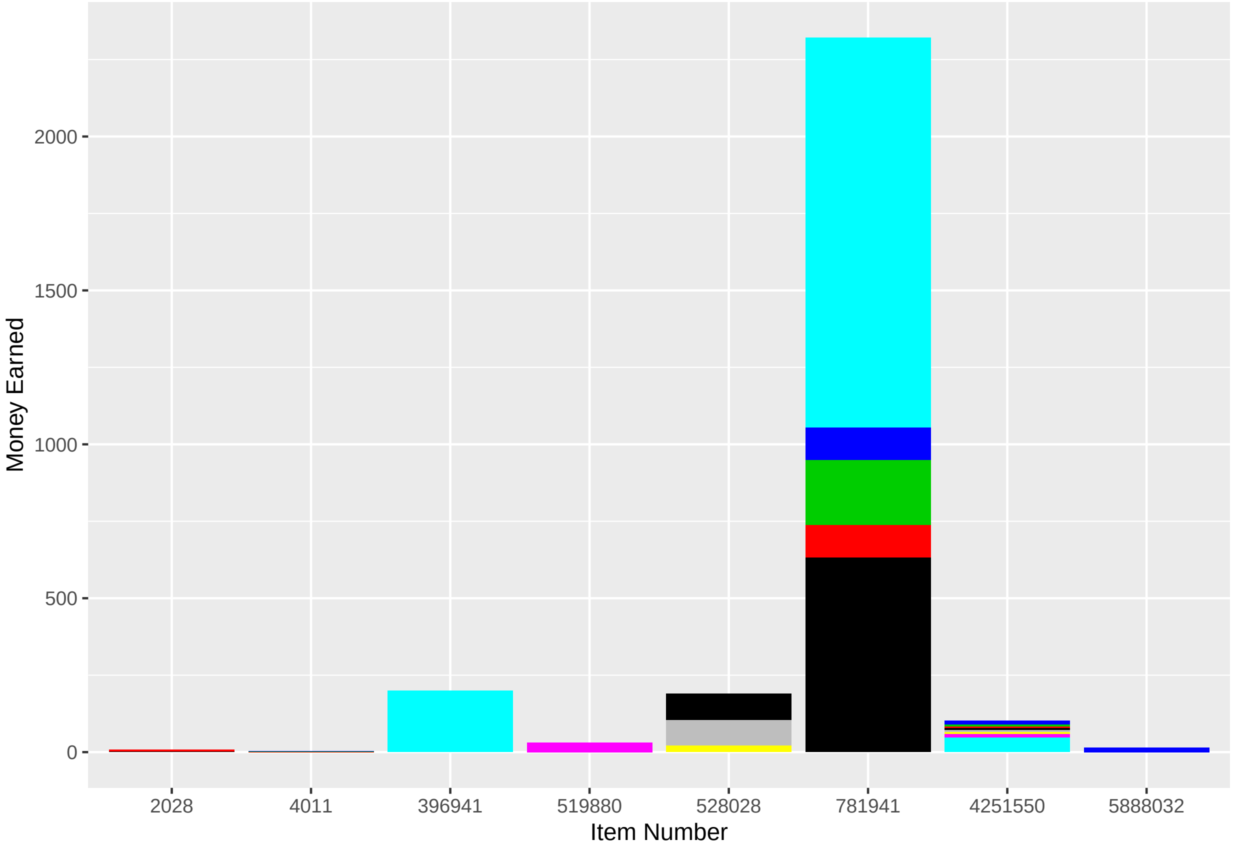
These two tables are showing roughly the same thing but in two different ways. One shows individual unit prices by showing how big the distance is between the start of one block height wise and where it ends. The other just shows the proportion of purchase ids, showing if it was a bulk purchase or many small purchases. Although now that I've turned it into a pdf it appears that the pdf also shows the unit prices.

```
ggplot(data = purchases, aes(item, ifelse(charge == "cost", price, 0), color = factor(pid))) + geom_col
```



```
ggplot(data = purchases, aes(item, ifelse(charge == "cost", price, 0))) + geom_col(fill = factor(purcha
```

Income
by Item Number


Money Earned
Item Number