Data Analysis - Office Sales and App Downloads

Elias Delosreyes

February 27, 2018

Contents

Load necessary packages and open xlsx

```r
#Load packages
library(tidyverse) #Used for data manipulation and visualization
library(readxl) #Used to read the excel file

#Open workbook
path <- 'Office Sales and App Downloads.xlsx'
```

Office Sales Analysis

1. Which product sub-category had the highest sales? How much sales did this sub-category have?

```r
#Load sheet and examine structure
orders <- read_excel(path, sheet = 'Orders')
orders <- data.frame(orders)

#Find Product Subcategory with most sales
orders %>%
  group_by(factor(`Product.Sub.Category`)) %>%
  summarize(Total_Sales = sum(Sales)) %>%
  arrange(desc(Total_Sales)) %>%
  head(5)
```

factor(Product.Sub.Category)	Total_Sales
Office Machines	318169.7
Chairs & Chairmats	261072.7
Telephones and Communication	198764.5
Tables	193764.6
Binders and Binder Accessories	185928.1

The office machines sub-category had the highest sales. The total sales were $318,169.68

2. What percent of total profit did the West region contribute?

```r
#Find percent of total profit
orders %>%
  mutate(Percentage_profit = Profit/sum(Profit)) %>%
  group_by(Region) %>%
  summarize(Total_Percent_Profit = sum(Percentage_profit) * 100) %>%
  arrange(desc(Total_Percent_Profit))
```

Region	Total_Percent_Profit
East	38.06333
Central	34.52619
West	33.84755
South	-6.43708

The West region contributed 33.85% of total profit

3. What is the averages sales per order for California?

```r
#Find average of sales in California
orders %>%
  filter(State.or.Province == 'California') %>%
```

```r
  group_by(`State.or.Province`) %>%
  summarize(mean(Sales))
```

State.or.Province	mean(Sales)
California	1347.246

The average sales per order in California is ~$1347.25.

4. Which product was ordered the most? How many times was it ordered?

```r
#Find most ordered item
orders %>%
  group_by(Product.Name) %>%
  summarize(Total.quantity = sum(Quantity.ordered.new)) %>%
  arrange(desc(Total.quantity)) %>%
  head(10)
```

Product.Name	Total.quantity
Newell 323	268
Economy Rollaway Files	216
Eldon Simplefile® Box Office®	183
Xerox 1923	159
Belkin 107-key enhanced keyboard, USB/PS/2 interface	154
Xerox 1922	150
Dixon Prang® Watercolor Pencils, 10-Color Set with Brush	146
Avery Hanging File Binders	139
Avery 493	137
Bevis 36 x 72 Conference Tables	136

The most ordered product was the Newell 323. This product was ordered 268 times.

App Downloads Analysis

5. How many downloads did each park have?

```r
#Load Downloads sheet and look at structure
downloads <- read_excel(path, 'Downloads')
downloads <- data.frame(downloads)
```

```r
downloads %>%
  group_by(Venue.ID) %>%
  summarize(Total.Downloads = sum(Downloads)) %>%
  arrange(desc(Total.Downloads))
```

Venue.ID	Total.Downloads
CF_CP	189655
CF_KBF	186690

Venue.ID	Total.Downloads
CF_KI	111647
CF_CW	84675
CF_CA	58102
CF_KD	36015
CF_GA	31391

- Cedar Point had 189,655 downloads.
- Knott's Berry Farm had 186,690 downloads.
- Kings Island had 111,647 downloads.
- Canada's Wonderland had 84,675 downloads.
- Carowinds had 58,102 downloads.
- Kings Dominion had 36,015 downloads.
- California's Great Adventure had 31,391 downloads.

6. How did downloads change month-over-month for Knott's Berry Farm?

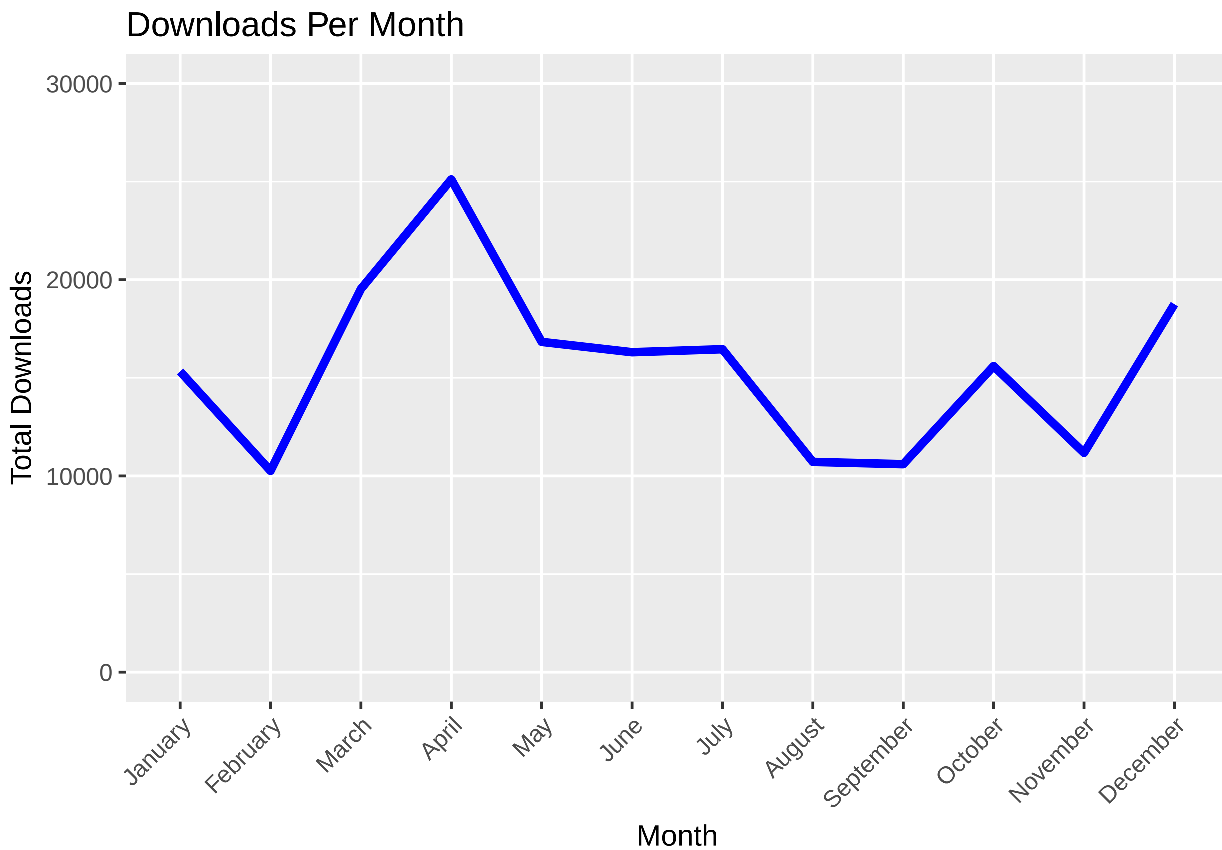
```
#Change month to factor variable
downloads$Month <- factor(downloads$Month, levels = c('January', 'February',
                                                      'March', 'April',
                                                      'May', 'June',
                                                      'July', 'August',
                                                      'September', 'October',
                                                      'November', 'December'))

#Table of month-over-month change in absolute value
Month.over.month <- downloads %>%
  filter(Venue.ID == 'CF_KBF') %>%
  group_by(Month) %>%
  summarize(Total.Downloads = sum(Downloads)) %>%
  arrange(Month)

#Table of month-over-month change in percentage
(Month.over.month <- Month.over.month %>%
  mutate(Abs_change = Total.Downloads - lag(Total.Downloads),
         Percent_change = Abs_change/lag(Total.Downloads) * 100))
```

Month	Total.Downloads	Abs_change	Percent_change
January	15328	NA	NA
February	10262	-5066	-33.050626
March	19532	9270	90.333268
April	25116	5584	28.588982
May	16834	-8282	-32.974996
June	16307	-527	-3.130569
July	16459	152	0.932115
August	10717	-5742	-34.886688
September	10600	-117	-1.091723
October	15599	4999	47.160377
November	11179	-4420	-28.335150
December	18757	7578	67.787816

```r
#Graph of month-over-month change
ggplot(Month.over.month, aes(x = Month, y = Total.Downloads, group = 1)) +
  geom_line(size = 1.5, color = 'blue') +
  ggtitle("Downloads Per Month") +
  ylim(0,30000) + ylab('Total Downloads') +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



The total month to month downloads have very high variance with ranges from 10,000 to 25,000 downloads within a given month. There is no obvious upward or downward trend in download growth.

7. What percent of downloads does each operating system (iOS vs Android) make up of the total downloads?

```r
#Create new `Platform` column and update it to iOS or Android
#depending on the name in the App column
downloads$Platform <- ifelse(str_detect(downloads$App, 'iOS') == 'TRUE', 'iOS', 'Android')

#Print Table of Downloads by Platform
downloads %>%
  mutate(Percent.Downloads = Downloads/sum(Downloads)) %>%
  group_by(Platform) %>%
  summarize(Sum.of.Downloads = sum(Downloads),
            Percent.of.Downloads = sum(Percent.Downloads) * 100) %>%
  arrange(desc(Sum.of.Downloads))
```

Platform	Sum.of.Downloads	Percent.of.Downloads
iOS	469632	67.26566
Android	228543	32.73434

There were more downloads on the iOS platform. iOS downloads accounted for 67.27% of downloads. Android downloads accounted for 32.73% of downloads.

8. What was the highest month for downloads?

```
#Find highest month for downloads
downloads %>%
  group_by(Month) %>%
  summarize(Sum.Downloads = sum(Downloads)) %>%
  arrange(desc(Sum.Downloads))
```

Month	Sum.Downloads
July	124129
June	105734
May	95629
August	95480
October	61418
April	56742
September	53881
March	29308
December	24919
January	20732
February	15910
November	14293

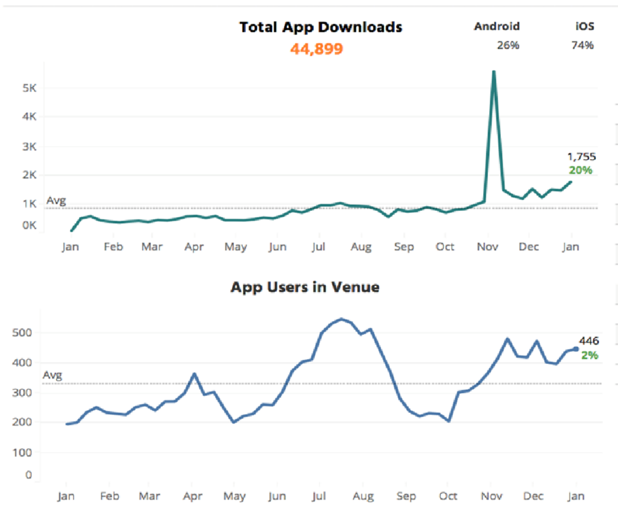
July was the month with the greatest number of downloads, with 124,129 during the month.

Promotion Analysis

9. During the month of November, there was a huge spike in downloads. What are some possible reasons for this?

During the month of November, downloads spiked to a little above 5,000. Based upon previous months, this 500% spike in growth was not natural, and some event probably triggered this. Some possible events could include:

- The app was not previously advertised to the customers of the resorts. If the resort did not advertise the mobile app until November, it could possibly explain the sudden spike and influx of users.
- There was a promotional offer for downloading the app. Promotional offers can incentivize users to download the app.
- The resort could have encouraged employees (employed at the resorts or even corporate office employees) to download the app during the month of November.



YTD Offer Details

Offer Name	# Redeem	$ Revenue	$ Cost	$ Profit	# Delivered	% Redeem
Samba Early Bird	1,559	$20,365	$9,211	$10,861	21,770	7.2%
Icecream 2x1	336	$1,972	$361	$1,546	7,022	4.8%
2x1 Pizza	140	$1,255	$370	$935	6,783	2.1%
Spatium 25% Off	40	$4,881	$2,009	$2,536	2,355	1.7%
Epazote Free App	29	$1,794	$39	$1,620	3,809	0.8%
Market Free App	15	$207	$66	$141	882	1.7%
La Cantina 20% Off	13	$826	$496	$295	3,905	0.3%
Coffee Offer	10	$56	$9	$46	3,560	0.3%
Ameca Early Bird	5	$226	$135	$91	3,809	0.1%
Free Wine	5	$268	$34	$247	1,201	0.4%
Epazote Free Margarita	4	$169	$6	$190	3,086	0.1%
Free Perfume	3	$331	$96	$235	741	0.4%
Ameca 10% Off Lunch	0	$0	$0	$0	577	0.0%
Grand Total	**2,159**	**$32,350**	**$12,832**	**$18,743**	**59,500**	**3.6%**

Figure 1: Picture references for questions 9-12

- On the side of extreme skepticism, the 500% spike in app downloads does not align with a 500% or even 100% spike for App users in the venue (unless the "App Users in Venue" graph is based upon a single resort, and the "Total App Downloads" graph is downloads across all resorts). However, if these two graphs are measuring the same downloads and app users for the same location(s), then it is possible that the spike in sales can be attributed to mass downloading by a botnet(network of bots) to arbitrarily inflate the app downloads and have it trend on the app stores.

10. For app downloads, if there were 1463 downloads for the week prior to the week with 1755 downloads, what does the 20% represents?

- The 20% represents the growth in downloads from the prior week. Using this formula we can verify that:

$$\frac{new\ value - old\ value}{old\ value} * (100) \quad = \quad \frac{1755 - 1463}{1463} * 100 \quad = \quad 20\%\ download\ growth$$

11. July and August were the months with the most users using the app in the resort. What are some possible reasons for this?

Reasons for spike in app venue users in the months of July and August:

- Summertime is the perfect time for vacation. For families with kids, summertime is the most opportune time due to long summer breaks for the kids. For those without kids, the summer is still one of the nicest times of the year in Mexico (where the resorts are located) due to the sunny weather.
- The resort may offer more promotions and/or advertise more during the summertime months of July and August.

12. What offers did well? Why?

- In terms of offer redemption rate, the Samba Early Bird, Icecream 2x1, and 2x1 Pizza deals had the highest redemption rate. Along with being the most redeemed, these three offers had very high return on investments:

```
Samba Early Bird: ~118% return on investment
Icecream 2x1: ~ 428% return on investment
2x1 Pizza: ~252% return on investment
```

  – All these deals are food related and thus appeal to a broad demographic.
  – Other promotional offers such as wine/alcohol, perfume, and Spa treatments do not necessarily appeal to all demographic groups, and thus have lower offer redemption rates.

- In terms of offers with the highest return on investment:

```
Epazote Free App: ~4150% return on investment
Epazote Free Margarita: ~3166% return on investment
Coffee Offer: ~511% return on investment
```

  – These offers do not have the highest redemption rates, but they do have some of the highest cost to profit ratios. The Epazote dining area could be a certain area of interest for the resort to promote further, as well as with coffee sales.

Additional Analysis

Are order priority and profit related?

```r
#Turn order priority from string variable to factor
orders$Order.Priority <- factor(orders$Order.Priority, levels = c('Not Specified',
    'Low',
    'Medium',
    'High',
    'Critical'))

#Table of order priority and average profits
orders %>%
  group_by(Order.Priority) %>%
  summarise(Average.Profits = mean(Profit), Number.orders = n()) %>%
  arrange(desc(Average.Profits))
```

Order.Priority	Average.Profits	Number.orders
Not Specified	177.99588	396
Medium	115.34607	376
Critical	97.96864	391
High	93.35174	391
Low	88.98204	398

There doesn't seem to be a clear relationship between profit and order priority. Orders with low priority have the lowest average profits and that seems to be correct. However, it is odd that orders with critical priority have less profits on average than orders with medium priority. Surprisingly, orders with no priority specified have the highest average profit.

```r
aov1 <- aov(Profit ~ Order.Priority, data = orders)
anova(aov1)
```

	Df	Sum Sq	Mean Sq	F value	Pr(>F)
Order.Priority	4	2137557	534389.4	0.4098975	0.8016323
Residuals	1947	2538332666	1303714.8	NA	NA

Lets run an statistical analysis of variance (ANOVA) to check if order priority is significant factor for profit.

- Technical Statistical Points:
  - The null hypothesis for our ANOVA is that order priority does not make a difference on profit levels. That is to say, the average profit between all order priority levels is equal to each other.
  - Given the F statistic, we fail to reject the null hypothesis at both the 5% and 10% significance level. We cannot reject the hypothesis that profit levels between all priority levels are equal to each other.

Our results suggest that: higher order priortity levels do not necessarily result in higher profits.

However, although the statistical ANOVA test says that order priority is not necessarily correlated with higher profits, it may still be false to assume that order priority and profit do not matter. If items are not delivered to clients that have "Critical" priority, then they may not want to partake in business again and the statistical model does not account for levels of customer loyalty. So, it is still very wise to pay attention to client priority levels.

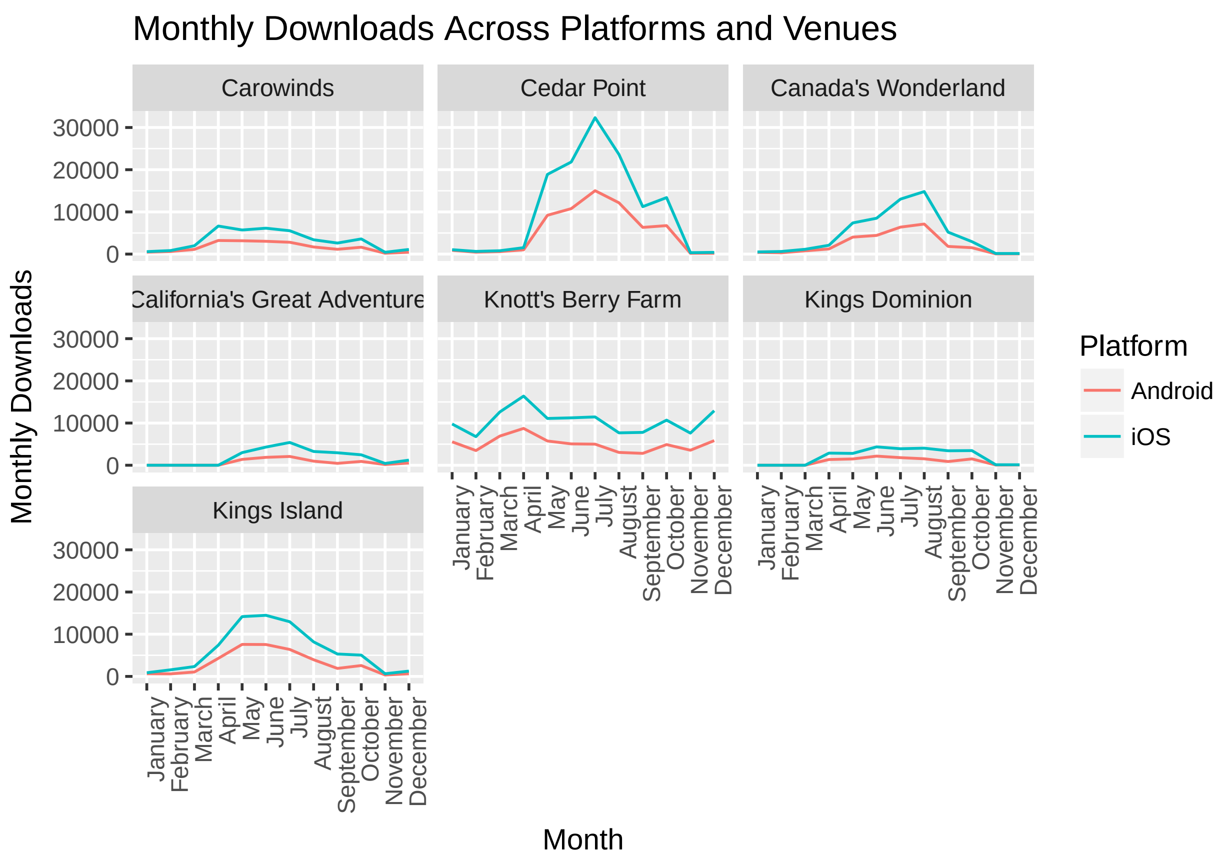
How has the growth in monthly downloads been across venues and platforms?

```r
#Group monthly downloads by venue, and platform
venue_month <- downloads %>%
  group_by(Month, Venue.ID, Platform) %>%
  summarize(Monthly_Download = sum(Downloads))

#Create functions to rename plot labels
Venue.Names <- list('CF_CA' = "Carowinds",
                    'CF_CP' = "Cedar Point",
                    'CF_CW' = "Canada's Wonderland",
                    'CF_GA' = "California's Great Adventure",
                    'CF_KBF' = "Knott's Berry Farm",
                    'CF_KD' = "Kings Dominion",
                    'CF_KI' = "Kings Island")

venue_labeller <- function(variable, value){
  return(Venue.Names[value])
}

#Plot monthly downloads by venue, and platform
ggplot(venue_month, aes(y = Monthly_Download, x = Month, col = Platform)) +
  geom_line(aes(group = Platform)) +
  facet_wrap(~ Venue.ID, labeller = venue_labeller) +
  theme(axis.text.x = element_text(angle = 90, hjust = 1)) +
  ylab('Monthly Downloads') +
  ggtitle('Monthly Downloads Across Platforms and Venues')
```



On average, iOS tends to outperform Android downloads, but that may be due to a higher population of iOS users rather than the quality of the apps on each respective platform. Cedar Point had very high downloads for a period of a couple months. Which leads to the question: Why did Cedar Point have so much more downloads for this extended period of time?

It could be that Cedar Point Parks are bigger than the others, and thus downloads are higher as a result. Or it could be Cedar Point's marketing tactics and advertising that lead to the spike in downloads. There are a number of possibilities for this reasoning, but more data would be required.

By knowing Cedar Point's tactics for inflating their app downloads, we could help other parks form strategies to increase their app downloads as well.

A good portion of items are being sold at a profit loss. What is the reasoning for this?

```
#Table of total profit, average profit, and
#number of orders amongst these categories (not to be confused
#with quantity purchased)
(loss_lead_table <- orders %>%
  group_by(Product.Sub.Category) %>%
  summarize(Total.profit = sum(Profit), Average.profit = mean(Profit),
            Total.orders = n()) %>%
  arrange(Average.profit) %>%
  filter(Average.profit < 0))
```

Product.Sub.Category	Total.profit	Average.profit	Total.orders
Tables	-7240.0714	-90.500892	80
Rubber Bands	-1544.8261	-45.436061	34
Scissors, Rulers and Trimmers	-1291.0959	-35.863775	36
Envelopes	-1194.4125	-21.716591	55
Bookcases	-930.4384	-21.638102	43
Pens & Art Supplies	-257.6288	-1.600179	161

```
#Print total lost profit
sum(loss_lead_table$Total.profit)
```

```
## [1] -12458.47
```

These item subcategories are all sold at a loss, and are called `loss leaders`. For example, we see tables sell for a big loss, but the same people who buy those tables typically buy more expensive chairs and chairmats. Within these subcategories, our total lost profit is -$12,458.47. Let's examine this further.

What do people typically tend to buy alongside these items that we sell at a loss?

```
#Create filtered data containing only the categories
#of loss leaders
loss_leaders <- orders %>%
  filter(Product.Sub.Category %in% c('Tables',
                                     'Rubber Bands',
                                     'Scissors, Rulers and Trimmers',
                                     'Envelopes', 'Bookcases',
                                     'Pens & Art Supplies')) %>%
  select(Product.Sub.Category, Customer.ID)

#Create filtered dataset to find customer ID's of people who buy
#loss leading items using left join
loss_lead_customers <- left_join(loss_leaders, orders, by = 'Customer.ID')

#Number of customers buying loss leading items
loss_lead_customers%>%
  distinct(Customer.ID) %>%
     summarize(Number.unique.customers = n())
```

Number.unique.customers
353

```
#Find the non-loss leading products these customers also purchase
(what_do_loss_leaders_buy <- loss_lead_customers %>%
  group_by(Product.Sub.Category.y) %>%
  summarize(Sum.profit = sum(Profit),
            Average.profit = mean(Profit),
            Total.orders = n()) %>%
  filter(!(Product.Sub.Category.y %in% c('Tables',
                                         'Rubber Bands',
```

```
                                          'Scissors, Rulers and Trimmers',
                                          'Envelopes', 'Bookcases',
                                          'Pens & Art Supplies'))
         & Average.profit > 0) %>%
  arrange(desc(Sum.profit)))
```

Product.Sub.Category.y	Sum.profit	Average.profit	Total.orders
Telephones and Communication	21532.806	331.27394	65
Chairs & Chairmats	15888.759	588.47257	27
Office Furnishings	13595.049	209.15459	65
Storage & Organization	12405.439	302.57168	41
Copiers and Fax	12108.895	1729.84219	7
Binders and Binder Accessories	6747.409	120.48945	56
Appliances	1643.445	63.20943	26
Labels	1641.475	65.65899	25

```
#Find total profit
sum(what_do_loss_leaders_buy$Sum.profit)
```

```
## [1] 85563.28
```

We see that there are 353 unique customer ID's of people that purchase these loss-leading categories. We profit off of things they buy in categories like telephones, chairs, office furnishings, and other various office materials. Our total profit in these categories from customers who also purchased loss-leading items is $85,563.28.

So even though we take a loss of -$12,458.47 by selling them tables, envelopes, bookcases, and other loss-leading items at lower than market value, we still come around $73,000 ahead when they decide to buy other things from us in their same orders.