Review Stats after TOS Change

```r
library(knitr)
library(ggplot2)
library(plotrix) # for pie3D
suppressMessages(library(dplyr))
```

Description of dataset

```r
#sum(dfRemovals$perc == 100)
#dfRemovals$perc == 100

df = bind_rows(
  read.delim("20161213--10-02.tab", header=T) %>% mutate(date="2016-10-02", after=F),
  read.delim("20161213--12-16.tab", header=T) %>% mutate(date="2016-12-16", after=T)
) %>% group_by(ASIN) %>% mutate(category = category[1]) %>% ungroup() %>% mutate(
  category = factor(ifelse(category == "Home and Kitchen", "Home & Kitchen", as.character(category)))
)
```

```
## Warning in bind_rows_(x, .id): Unequal factor levels: coercing to character
```

```r
calcDelta = function(df) {
  if (sum(complete.cases(df)) != 2)
    return (data.frame())
  with(df, data.frame(category = category[1]) %>% mutate(
    category = category,
    reviews0 = Total[1], reviews1 = Total[2], reviewsDiff = reviews1 - reviews0,
    rating0 = Avg[1], rating1 = Avg[2], ratingDiff = rating1 - rating0
  ))
}
dfDelta = df %>% group_by(ASIN) %>% do(calcDelta(.)) %>% ungroup()
dfLoss = dfDelta %>% filter(reviewsDiff < 0) %>% mutate(reviewLossPerc = -100 * reviewsDiff / reviews0,
#dfLoss
data.frame(nOriginal = length(unique(df$ASIN)), nNonEmpty = nrow(dfDelta), nLoss = nrow(dfLoss))
```

```
##   nOriginal nNonEmpty nLoss
## 1      7967      5995  1548
```

Info:

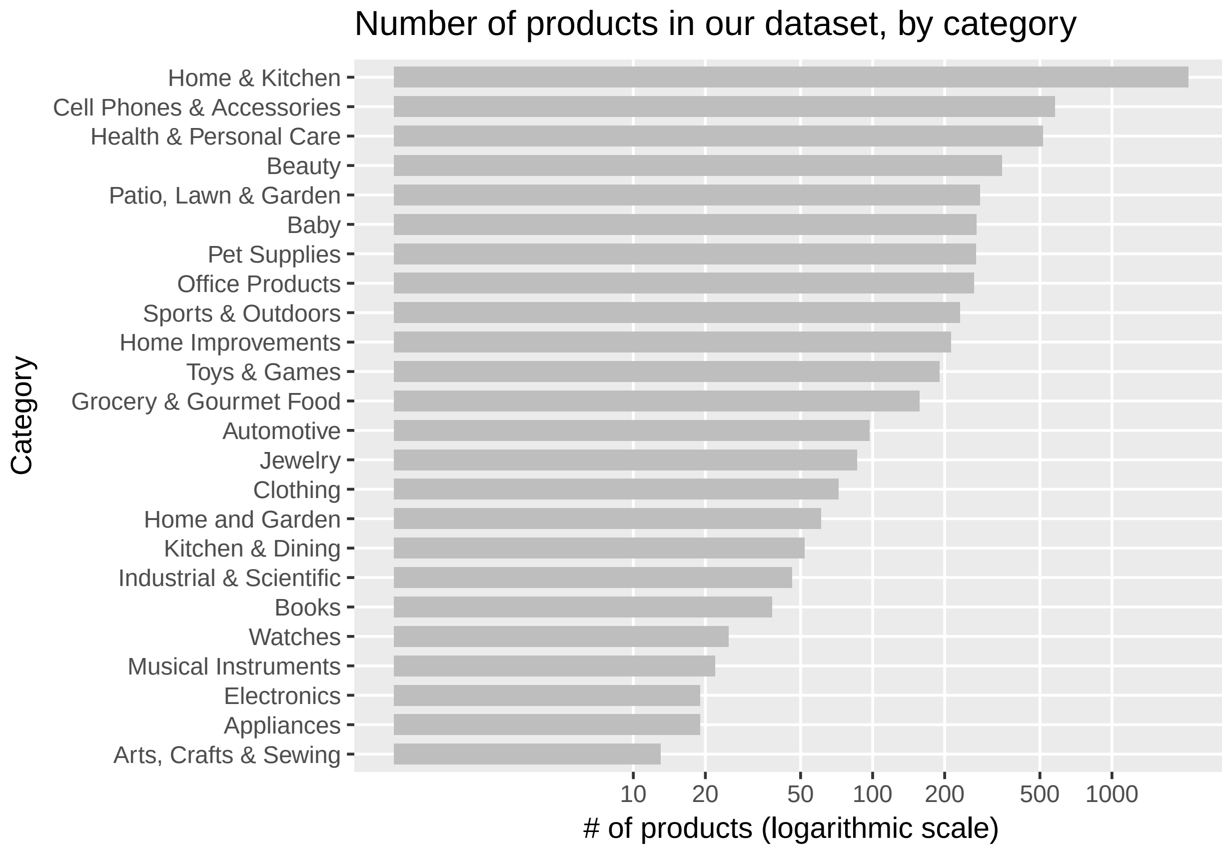
- We followed 7967 products. Of those, we have good data for 5995 both before and after the TOS changes. Of those, 1548 lost reviews.

```r
slices = c(nrow(dfDelta) - nrow(dfLoss), nrow(dfLoss))
lbls = c(paste0("No Loss (", nrow(dfDelta) - nrow(dfLoss), ")"), paste0("Loss (", nrow(dfLoss), ")"))
pie3D(slices,labels=lbls,explode=0.1, main="How many products lost reviews?", theta=1, radius=0.7, col=
```


```
dfCategoryProportions = dfDelta %>% group_by(category) %>% summarize(n = n(), loss = sum(reviewsDiff < 0
dfCategoryProportions
```

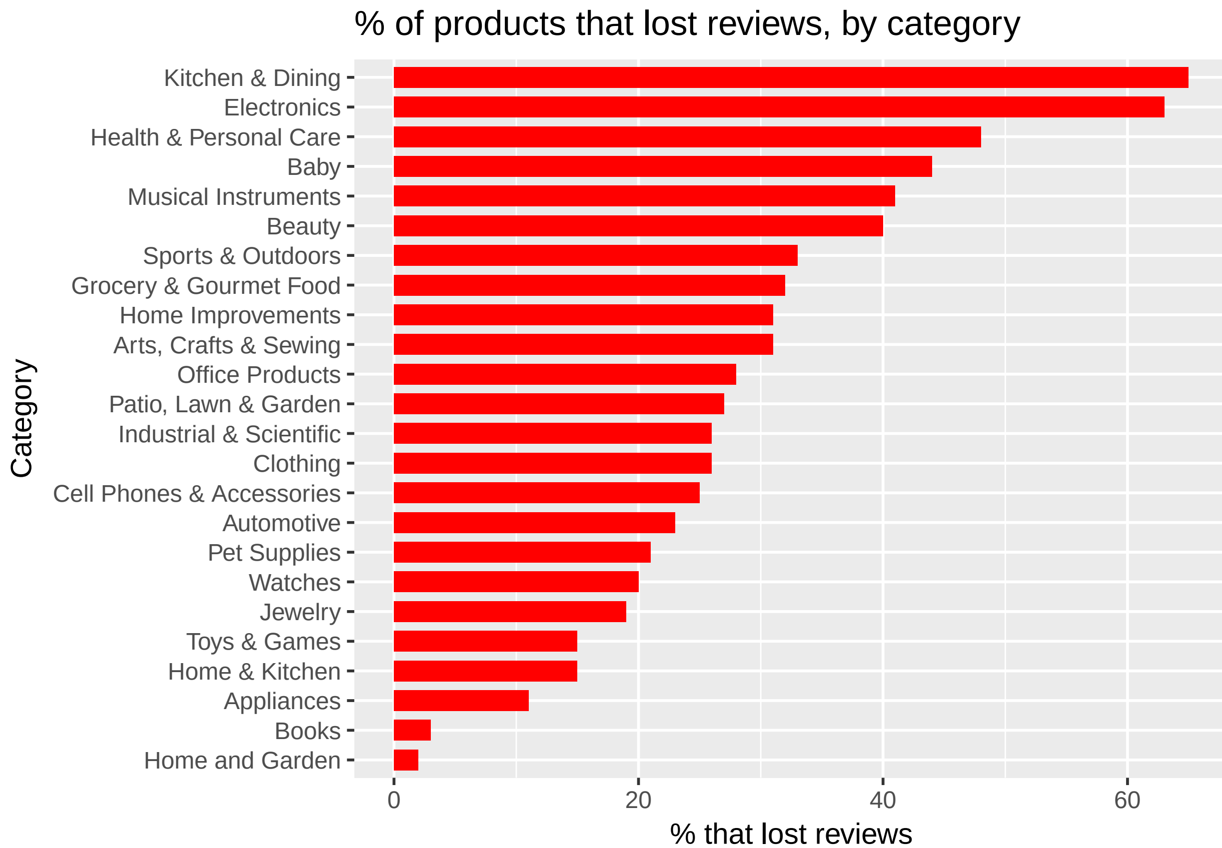
```
## # A tibble: 24 × 5
##                    category     n  loss noloss lossPerc
##                      <fctr> <int> <int>  <int>    <dbl>
## 1           Home and Garden    61     1     60        2
## 2                     Books    38     1     37        3
## 3                Appliances    19     2     17       11
## 4             Home & Kitchen  2092   319   1773       15
## 5              Toys & Games   190    29    161       15
## 6                   Jewelry    86    16     70       19
## 7                   Watches    25     5     20       20
## 8              Pet Supplies   270    56    214       21
## 9                Automotive    97    22     75       23
## 10 Cell Phones & Accessories   578   143    435       25
## # ... with 14 more rows
```

```
dfTemp = dfCategoryProportions %>% arrange(n)
dfTemp$category = factor(dfTemp$category, levels=dfTemp$category)
ggplot(dfTemp, aes(x=category, y=n)) +
  geom_bar(stat = "identity", width = 0.7, fill="grey") +
  scale_y_log10(breaks=c(10, 20, 50, 100, 200, 500, 1000), minor_breaks=NULL) +
  coord_flip() +
  labs(title="Number of products in our dataset, by category", x="Category", y="# of products (logarithm
```



In the graph above, we see the number of products we followed in each category (but only for categories with at least 10 products).

```
dfTemp = dfCategoryProportions
dfTemp$category = factor(dfTemp$category, levels=dfTemp$category)
ggplot(dfTemp, aes(x=category, y=lossPerc)) +
  geom_bar(stat = "identity", width = 0.7, fill="red") +
  coord_flip() +
  labs(title="% of products that lost reviews, by category", x="Category", y="% that lost reviews")
```



Question 1

Question: On average, how many reviews were removed after 10/3/16? As a total number, and as a percentage of existing reviews, per product.

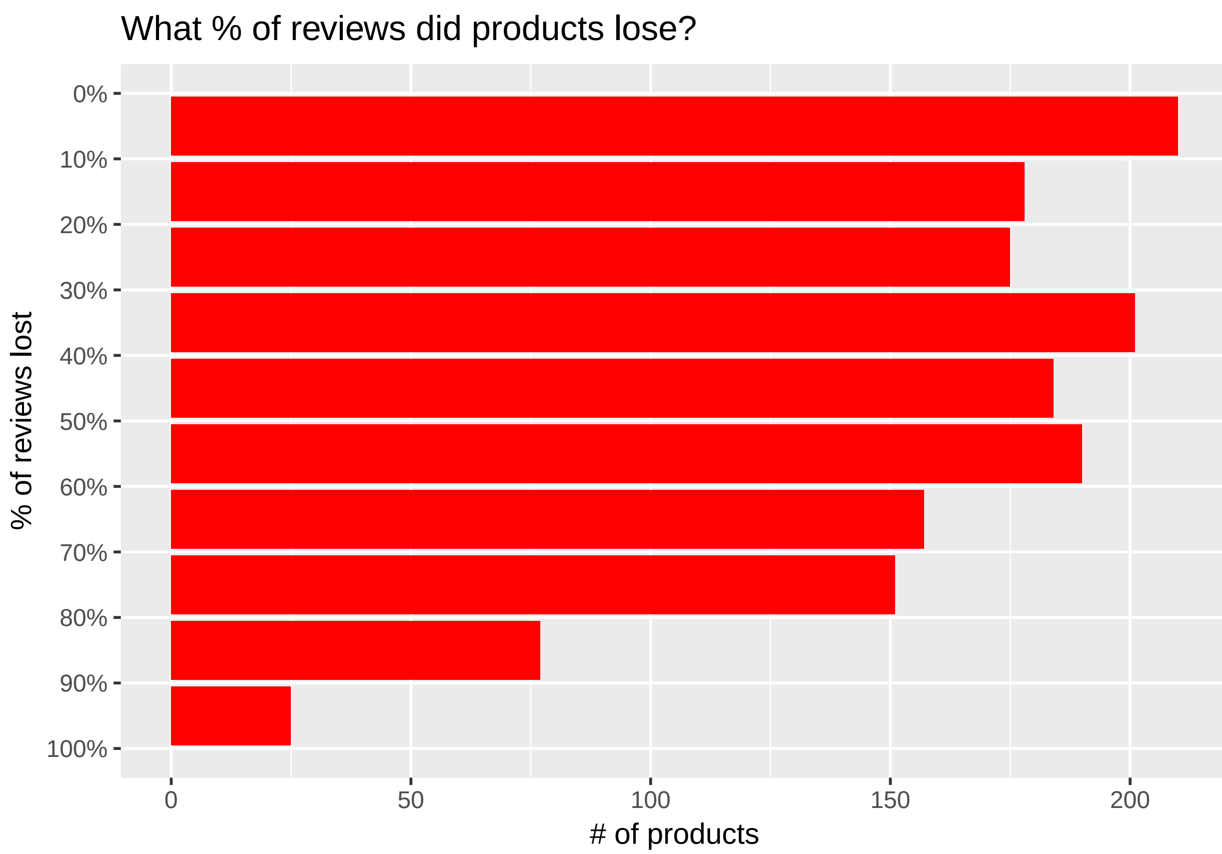
```
ggplot(dfLoss %>% mutate(loss = -reviewsDiff), aes(x=loss, y=reviewLossPerc)) + geom_point(color="red")
```

```
## `geom_smooth()` using method = 'gam'
```

In the graph above, each product that lost reviews is shown as a dot. On the x-axis are the number of reviews the product lost, and on the y-axis is the percentage of reviews that were lost. We can see along the top of the graph that 9 products lost 100% of their reviews, and on the right side of the graph we see that 5 products lost more than 2000 reviews. The blue trend line highlights that products which lost more reviews tended to lose a higher percentage or their reviews, and if a product lost more than 200 reviews, that tended to represent about 2/3rds of their reviews on average.

```
x = dfLoss %>% mutate(group = ordered(11 - .bincode(reviewLossPerc, breaks=c(0, 10, 20, 30, 40, 50, 60,
ggplot(x, aes(x=percMid, y=count)) + geom_histogram(stat="identity", fill="red") + scale_x_continuous(b
```

```
## Warning: Ignoring unknown parameters: binwidth, bins, pad
```



In the graph above, we see the percentage of reviews that products lost. For example, 210 products lost a maximum of 10% or their reviews, and 25 products losts between 90% and 100%.

```
median(dfLoss$reviewsDiff)
```

```
## [1] -28
```

```
median(dfLoss$reviewLossPerc)
```

```
## [1] 40.64171
```

Among products that lost reviews, their median loss was 28 reviews or 41% of their reviews.

Question 2

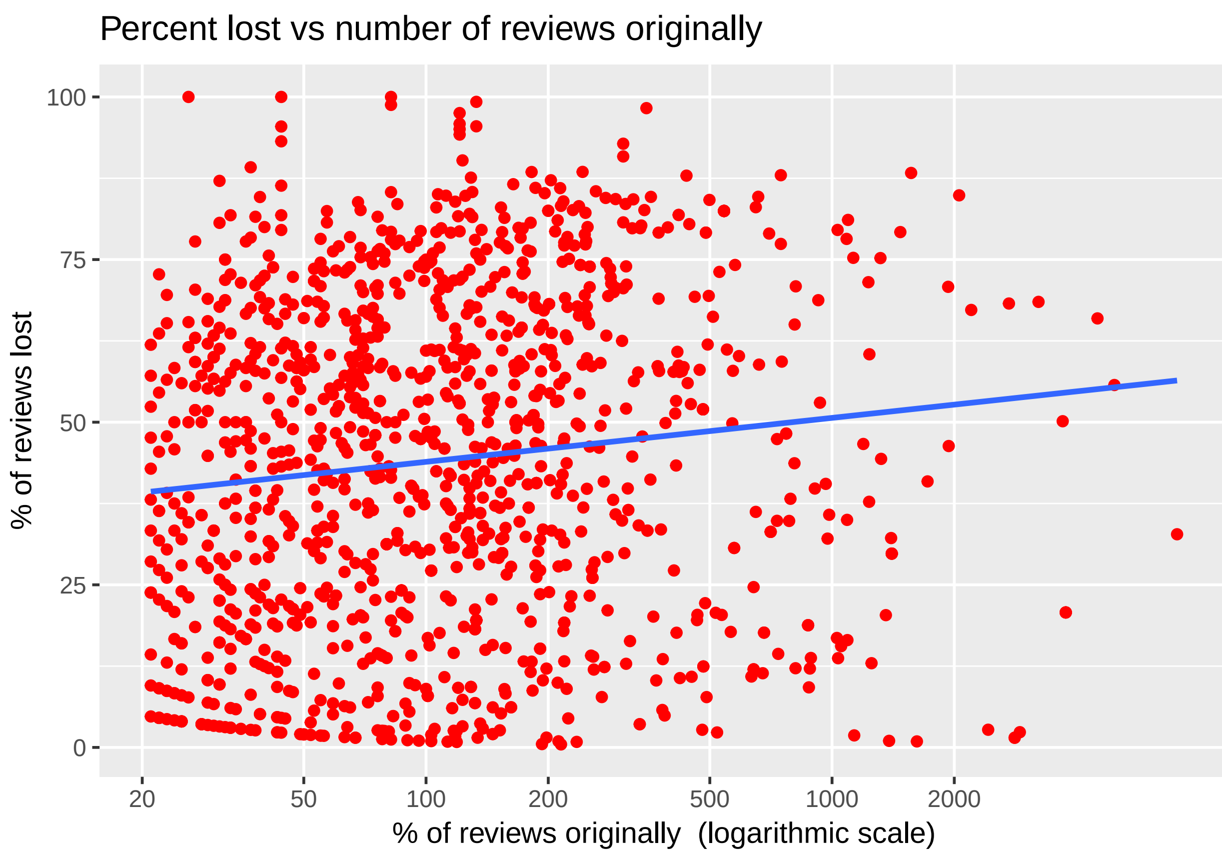
Question: Were any types of sellers targeted in particular? Sellers in a certain category, larger/smaller sellers by revenue/SKU, products at a certain price range?

```
lm1 = lm(data=dfLoss, reviewLossPerc ~ log(reviews0) + category)
summary(lm1)
```

```
##
## Call:
## lm(formula = reviewLossPerc ~ log(reviews0) + category, data = dfLoss)
##
## Residuals:
##     Min      1Q  Median      3Q     Max
## -59.376 -19.295  -0.631  19.022  75.358
```

```
##
## Coefficients:
##                                 Estimate Std. Error t value Pr(>|t|)
## (Intercept)                      18.5320    17.2557   1.074    0.283
## log(reviews0)                     3.6172     0.5037   7.181 1.08e-12 ***
## categoryArts, Crafts & Sewing    15.5023    20.9911   0.739    0.460
## categoryAutomotive               -0.6135    17.9018  -0.034    0.973
## categoryBaby                      8.9999    17.2831   0.521    0.603
## categoryBeauty                   15.0438    17.2683   0.871    0.384
## categoryBooks                   -30.4506    29.6859  -1.026    0.305
## categoryCamera & Photo           24.8276    22.1289   1.122    0.262
## categoryCell Phones & Accessories  1.8313  17.2621   0.106    0.916
## categoryClothing                 18.7968    18.0205   1.043    0.297
## categoryElectronics               8.6348    18.5138   0.466    0.641
## categoryGrocery & Gourmet Food   -1.4031    17.4761  -0.080    0.936
## categoryHealth & Personal Care   12.8888    17.2111   0.749    0.454
## categoryHome & Kitchen            3.6104    17.1928   0.210    0.834
## categoryHome and Garden          31.3309    29.6891   1.055    0.291
## categoryHome Improvements         6.1105    17.4008   0.351    0.726
## categoryHome Theater             30.0432    29.6914   1.012    0.312
## categoryIndustrial & Scientific  -1.5512    18.5123  -0.084    0.933
## categoryJewelry                  -0.2392    18.1799  -0.013    0.990
## categoryKitchen & Dining         15.0986    17.6394   0.856    0.392
## categoryLaunchpad                -9.1571    29.6911  -0.308    0.758
## categoryMusic                     3.0006    29.7034   0.101    0.920
## categoryMusical Instruments       5.9816    18.9480   0.316    0.752
## categoryOffice Products           5.2557    17.3724   0.303    0.762
## categoryPatio, Lawn & Garden      5.9715    17.3649   0.344    0.731
## categoryPet Supplies              7.1049    17.4426   0.407    0.684
## categoryPrime Pantry            -45.0273    29.7540  -1.513    0.130
## categoryShoes                    18.3391    29.6926   0.618    0.537
## categorySports & Outdoors        11.0613    17.3632   0.637    0.524
## categoryToys & Games              8.9879    17.7205   0.507    0.612
## categoryVideo Games              33.1827    19.4365   1.707    0.088 .
## categoryWatches                  18.5711    20.2845   0.916    0.360
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 24.24 on 1516 degrees of freedom
## Multiple R-squared:  0.0895, Adjusted R-squared:  0.07088
## F-statistic: 4.807 on 31 and 1516 DF,  p-value: < 2.2e-16
```

```r
ggplot(dfLoss %>% filter(reviews0 > 20), aes(x=reviews0, y=reviewLossPerc)) + geom_point(color="red") +
```



As the blue trend line shows in graph above, products that had a lot of reviews before the TOS change tended to lose a slightly higher percentage their reviews. However, the difference was not very large. No particular categories were targeted for removal (this is shown in the table above, but this might be too technical to try to explain in a blog post).

Question 3

Question: Was there an overall change in rating for listings that lost reviews? Positive or negative, and by how much?

```
median(dfLoss$ratingDiff)
```

```
## [1] -0.13
```

```
median(dfLoss$ratingLossPerc)
```

```
## [1] 3.117773
```

Among products that lost reviews, their median loss in star rating was just 0.13, or about 3% of their star rating.