FinalProject-QuantitativeAnalysis

```r
library(tidyverse) # data viz
library(tidytext) # text mining
library(gridExtra) # for plotting side by side
library(scales) # for previewing colors
```

1) Load in Data

```r
all <- read_csv("all_products.csv")
dryer <- read_csv("hair_dryer.csv")
paci <- read_csv("pacifier.csv")
micro <- read_csv("microwave.csv")
# head(dryer)
# head(paci)
# head(micro)
```

2) Preview Columns

Function to preview columns

```r
product_summary <- function(df){
  print(deparse(substitute(df)))
  cat("marketplace: ", unique(df$marketplace),"\n")
  cat("category: ", unique(df$product_category),"\n")
  cat("# of titles: ",length(unique(df$product_title)),"\n")
  cat("# of ids: ", length(unique(df$product_id)),"\n")
  cat("# of customers: ",length(unique(df$customer_id)),"\n")
  cat("# of reviews: ",length(unique(df$review_id)),"\n")
  cat("# of parents: ", length(unique(df$product_parent)),"\n")
  print(summary(df$star_rating))
  print(summary(df$review_date))
  cat("\n")
}
```

```r
product_summary(dryer)
```

```
## [1] "dryer"
## marketplace:  us
## category:  Beauty
## # of titles:  503
## # of ids:  538
## # of customers:  11348
## # of reviews:  11470
## # of parents:  473
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   4.000   5.000   4.116   5.000   5.000
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "2002-03-02" "2012-12-20" "2014-03-19" "2013-07-26" "2015-01-17" "2015-08-31"
```

```
product_summary(paci)
```

```
## [1] "paci"
## marketplace:  us
## category:  baby
## # of titles:  5533
## # of ids:  6482
## # of customers:  17661
## # of reviews:  18939
## # of parents:  5432
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   4.000   5.000   4.305   5.000   5.000
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "2003-04-27" "2013-05-30" "2014-07-28" "2014-01-09" "2015-02-10" "2015-08-31"
```

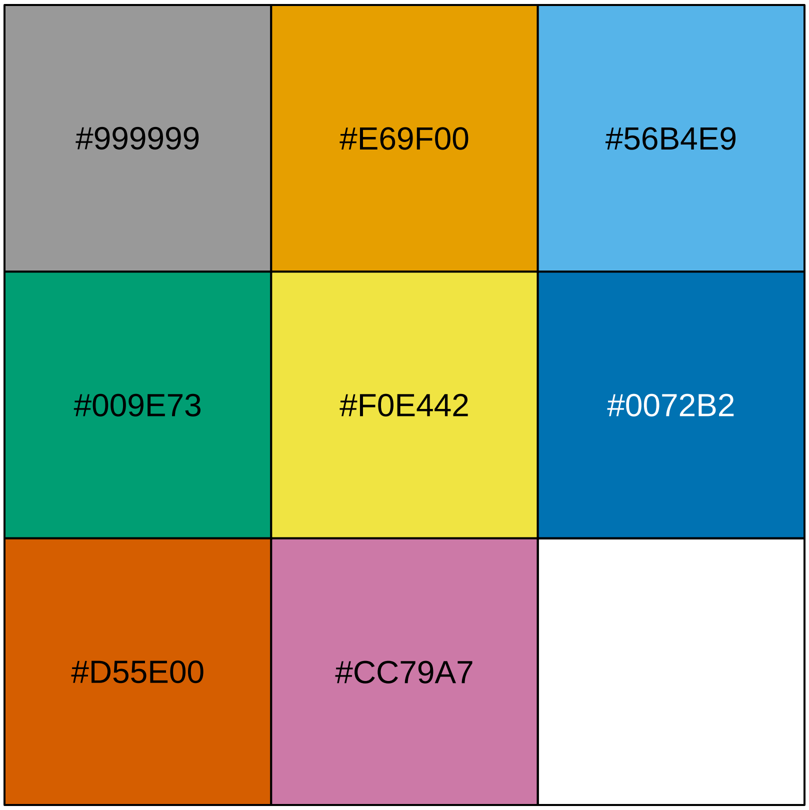
```
product_summary(micro)
```

```
## [1] "micro"
## marketplace:  us
## category:  major appliances
## # of titles:  43
## # of ids:  56
## # of customers:  502
## # of reviews:  502
## # of parents:  42
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   1.000   3.000   4.000   3.743   5.000   5.000
##         Min.      1st Qu.       Median         Mean      3rd Qu.         Max.
## "2015-01-22" "2015-03-09" "2015-05-16" "2015-05-13" "2015-07-16" "2015-08-31"
##         NA's
##          "1"
```

3) Look at distributions of star ratings, helpful votes, total votes, vine, verified, and dates

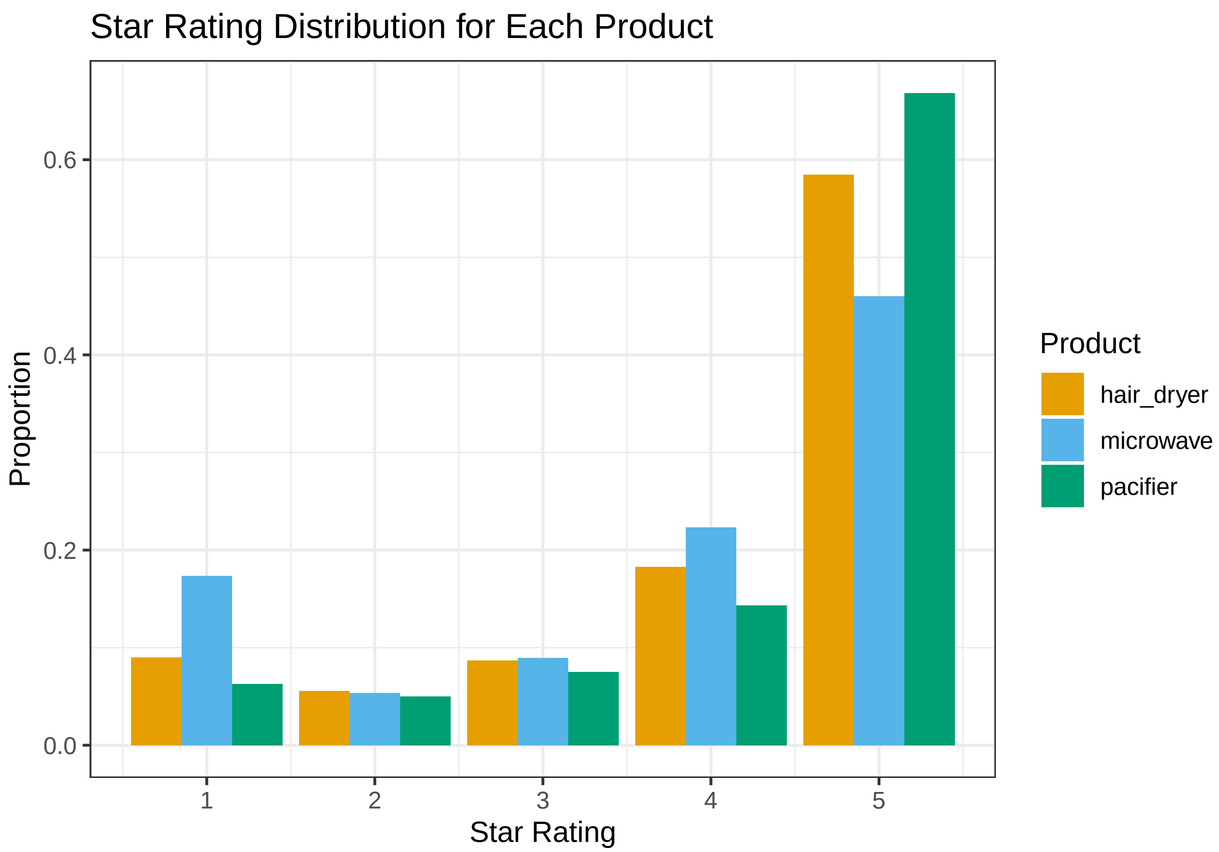
Define custom color set

```
cbp <- c("#999999", "#E69F00", "#56B4E9", "#009E73",
          "#F0E442", "#0072B2", "#D55E00", "#CC79A7")
show_col(cbp)
```



Star Ratings

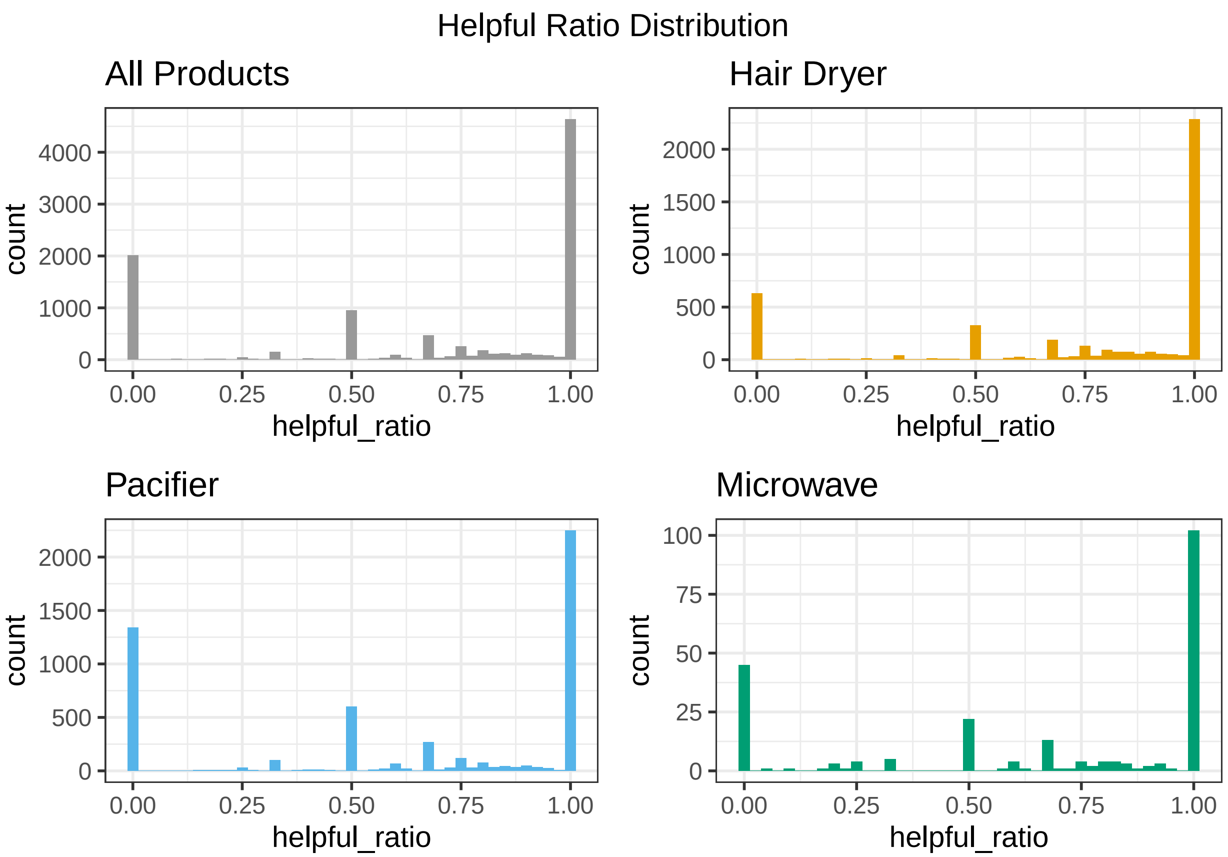
```
# create summary data for star ratings
star_all <- all %>% group_by(type) %>% summarise(rating_ct_all = n())
star_type <- all %>% group_by(type,star_rating) %>% summarise(rating_ct = n())
star_type<- inner_join(star_all,star_type,by="type")
star_type <- star_type %>% mutate(rating_p = rating_ct/rating_ct_all)
# print(star_type)
rate_hist <- ggplot(star_type,aes(x=star_rating,y=rating_p,fill=type)) + geom_bar(stat="identity",positi
  ggtitle("Star Rating Distribution for Each Product") + ylab("Proportion") + xlab("Star Rating") + labs
  scale_fill_manual(values=cbp[c(2:4)]) + theme_bw()
print(rate_hist)
```



Microwaves seem to have the worst ratings (highest chance of malfunction?), however it does slightly lead in 3 and 4 star ratings. Pacifier has best 5 star ratings, probably since least likely to malfunction. Hairdryer has similar trend as pacifier.

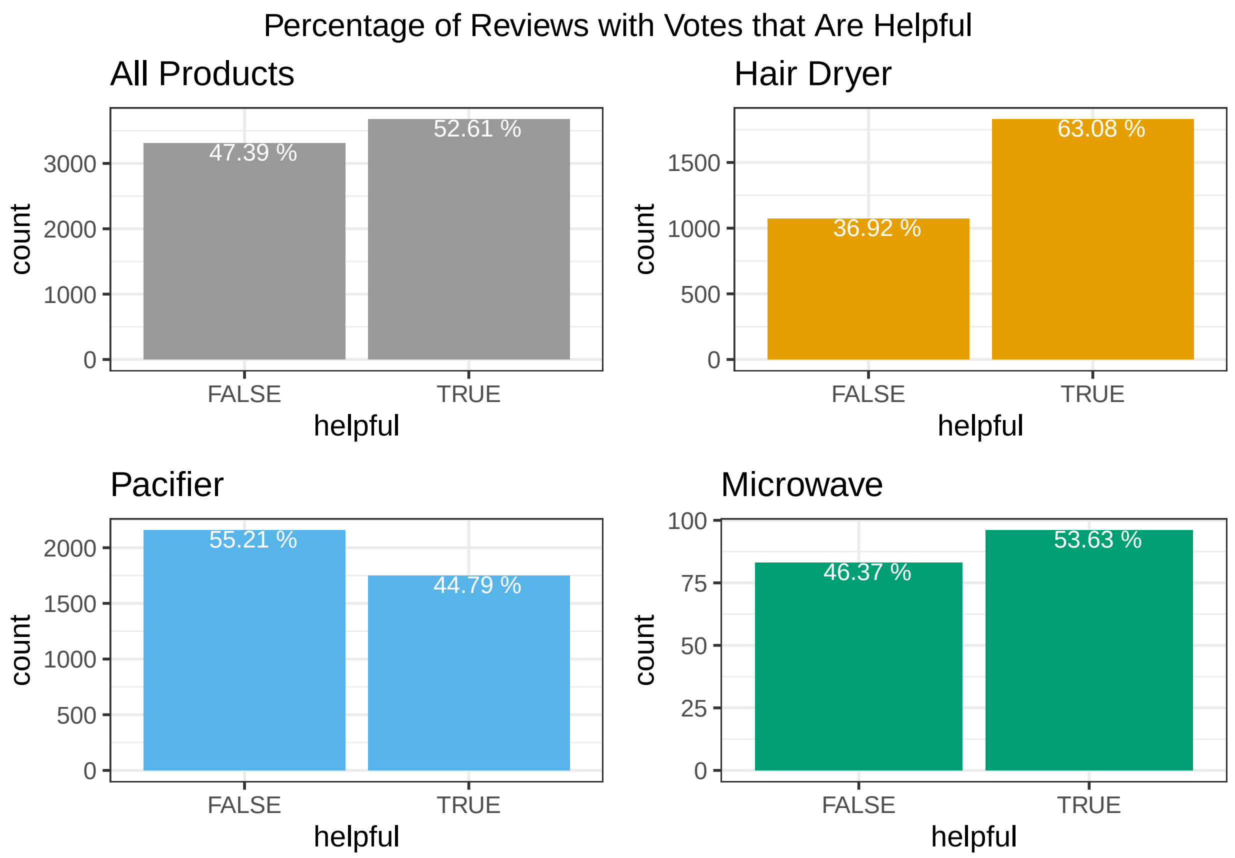
look at helpful ratio

```
hr1 <- ggplot(all,aes(x=helpful_ratio)) + geom_histogram(binwidth = 0.025,fill=cbp[1]) +
  theme(legend.position = "none") + theme_bw() + ggtitle("All Products")
hr2 <- ggplot(dryer,aes(x=helpful_ratio)) + geom_histogram(binwidth = 0.025,fill=cbp[2]) +
  theme(legend.position = "none") + theme_bw() + ggtitle("Hair Dryer")
hr3 <- ggplot(paci,aes(x=helpful_ratio)) + geom_histogram(binwidth = 0.025,fill=cbp[3]) +
  theme(legend.position = "none") + theme_bw() + ggtitle("Pacifier")
hr4 <- ggplot(micro,aes(x=helpful_ratio)) + geom_histogram(binwidth = 0.025,fill=cbp[4]) +
  theme(legend.position = "none") + theme_bw() + ggtitle("Microwave")
grid.arrange(hr1,hr2,hr3,hr4, ncol=2, top = "Helpful Ratio Distribution")
```



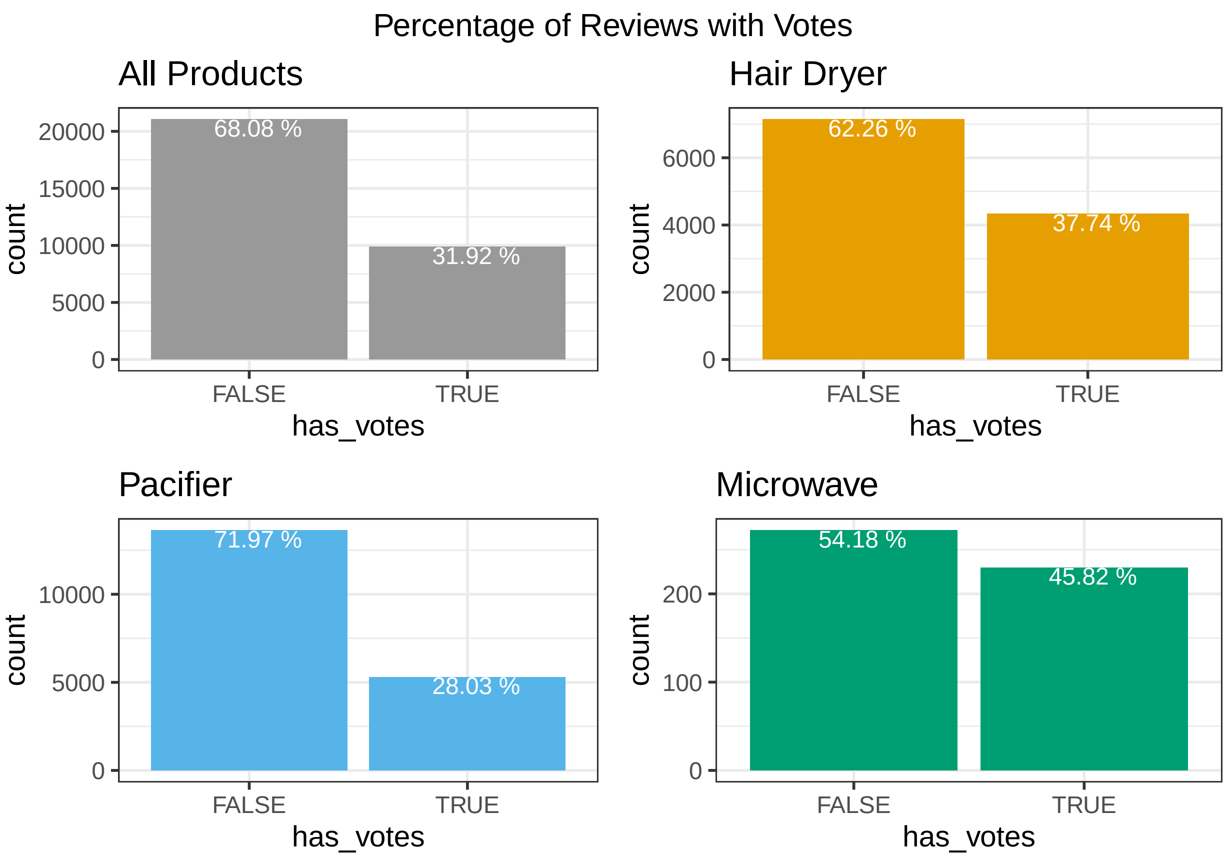
look at helpful indicator

```
h1 <- ggplot(data=subset(all, !is.na(helpful)),aes(x=helpful)) + geom_bar(fill=cbp[1]) +
  geom_text(stat='count', aes(label=paste(round(..count../sum(!is.na(all$helpful)) * 100,2),"%")),
            hjust=0.4, vjust=1, size = 3,color="white") +
  theme(legend.position = "none") + theme_bw() + ggtitle("All Products")
h2 <- ggplot(data=subset(dryer, !is.na(helpful)),aes(x=helpful)) + geom_bar(fill=cbp[2]) +
  geom_text(stat='count', aes(label=paste(round(..count../sum(!is.na(dryer$helpful)) * 100,2),"%")),
            hjust=0.4, vjust=1, size = 3,color="white") +
  theme(legend.position = "none") + theme_bw() + ggtitle("Hair Dryer")
h3 <- ggplot(data=subset(paci, !is.na(helpful)),aes(x=helpful)) + geom_bar(fill=cbp[3]) +
  geom_text(stat='count', aes(label=paste(round(..count../sum(!is.na(paci$helpful)) * 100,2),"%")),
            hjust=0.4, vjust=1, size = 3,color="white") +
  theme(legend.position = "none") + theme_bw() + ggtitle("Pacifier")
h4 <- ggplot(data=subset(micro, !is.na(helpful)),aes(x=helpful)) + geom_bar(fill=cbp[4]) +
  geom_text(stat='count', aes(label=paste(round(..count../sum(!is.na(micro$helpful)) * 100,2),"%")),
            hjust=0.4, vjust=1, size = 3,color="white") +
  theme(legend.position = "none") + theme_bw() + ggtitle("Microwave")
grid.arrange(h1,h2,h3,h4, ncol=2, top = "Percentage of Reviews with Votes that Are Helpful")
```



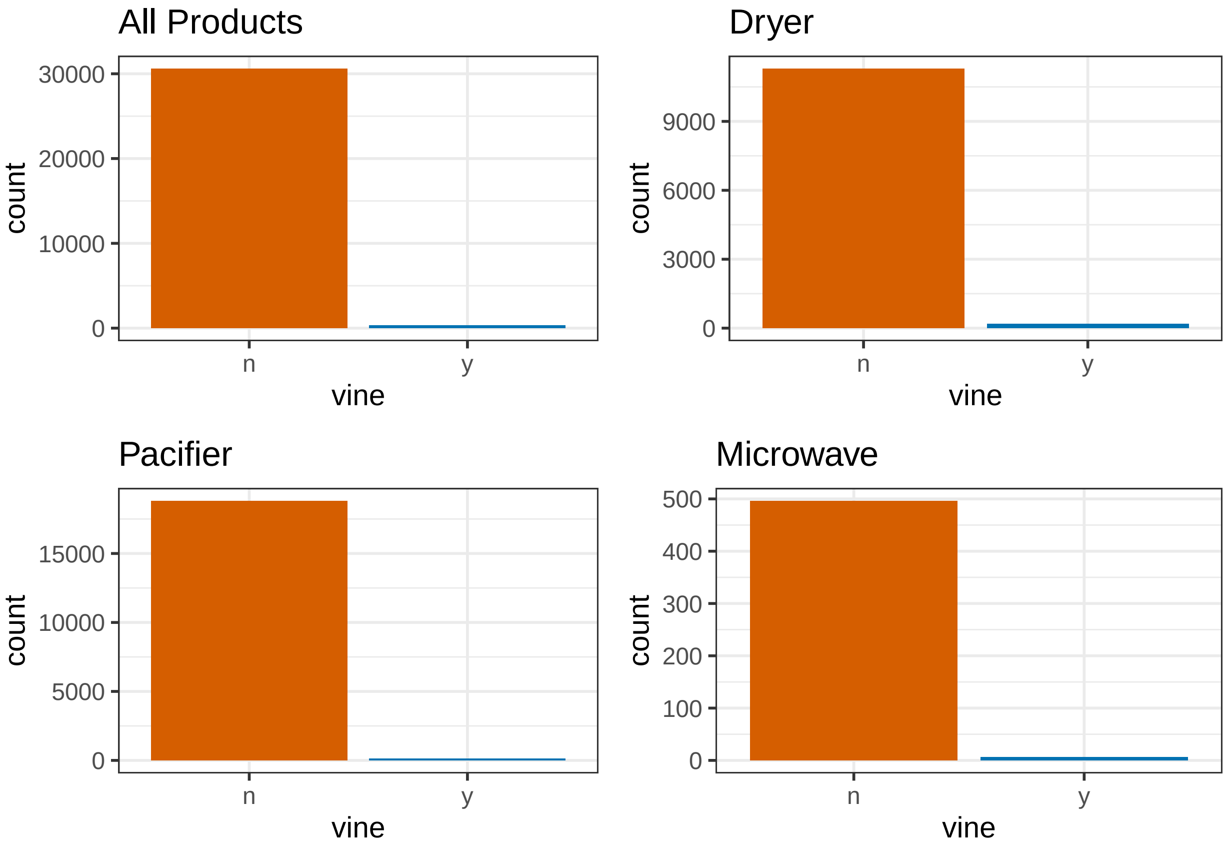
look at proportion of reviews that have any votes (total_votes)

```
hv1 <- ggplot(all,aes(x=has_votes)) + geom_bar(fill=cbp[1]) +
  theme(legend.position = "none") + theme_bw() +
  geom_text(stat='count', aes(label=paste(round(..count../nrow(all) * 100,2),"%")),
            hjust=0.4, vjust=1, size = 3,color="white") +
  theme(legend.position = "none") + theme_bw() + ggtitle("All Products")

hv2 <- ggplot(dryer,aes(x=has_votes)) + geom_bar(fill=cbp[2]) +
  theme(legend.position = "none") + theme_bw() +
  geom_text(stat='count', aes(label=paste(round(..count../nrow(dryer) * 100,2),"%")),
            hjust=0.4, vjust=1, size = 3,color="white") +
  theme(legend.position = "none") + theme_bw() + ggtitle("Hair Dryer")

hv3 <- ggplot(paci,aes(x=has_votes)) + geom_bar(fill=cbp[3]) +
  theme(legend.position = "none") + theme_bw() +
  geom_text(stat='count', aes(label=paste(round(..count../nrow(paci) * 100,2),"%")),
            hjust=0.4, vjust=1, size = 3,color="white") +
  theme(legend.position = "none") + theme_bw() + ggtitle("Pacifier")

hv4 <- ggplot(micro,aes(x=has_votes)) + geom_bar(fill=cbp[4]) +
  theme(legend.position = "none") + theme_bw() +
  geom_text(stat='count', aes(label=paste(round(..count../nrow(micro) * 100,2),"%")),
            hjust=0.4, vjust=1, size = 3,color="white") +
  theme(legend.position = "none") + theme_bw() + ggtitle("Microwave")
grid.arrange(hv1,hv2,hv3,hv4, ncol=2, top = "Percentage of Reviews with Votes")
```



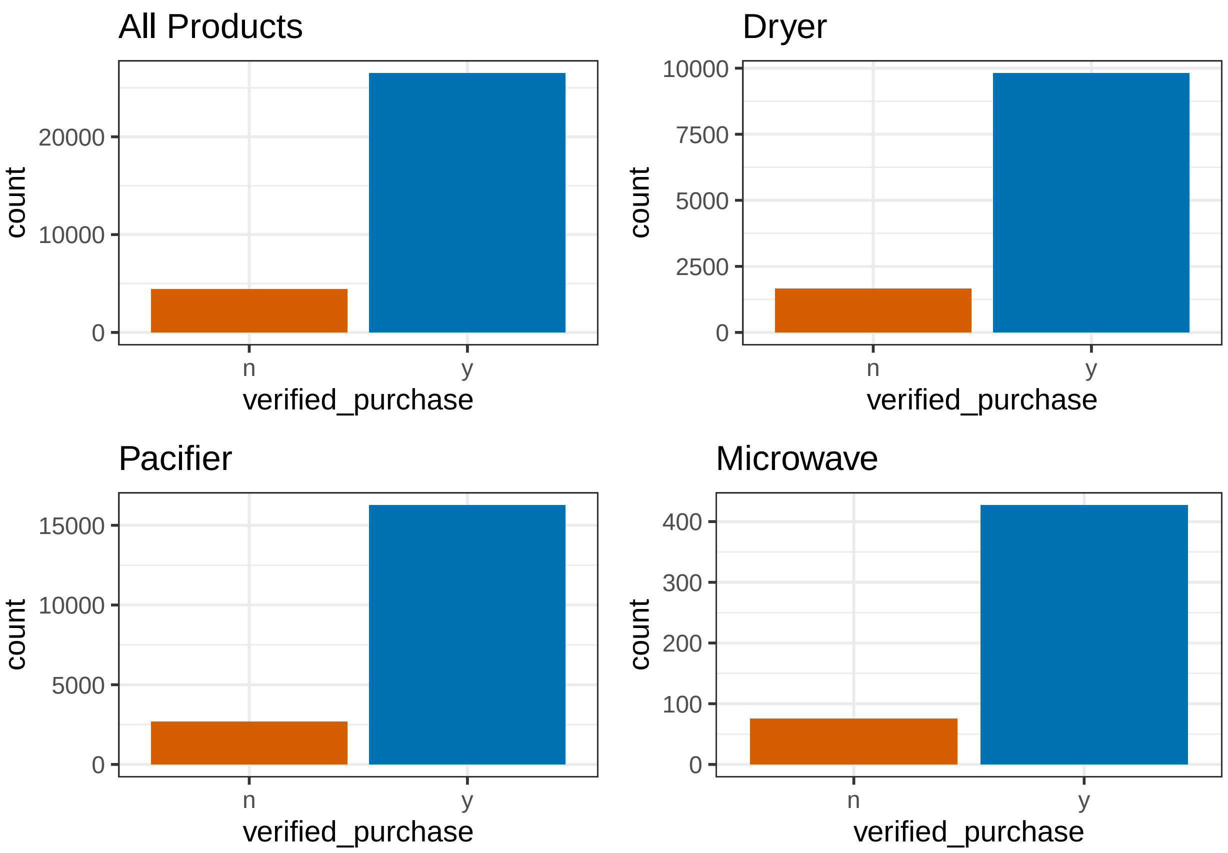
look at vine, nonvine ratio

```
v1 <- ggplot(all,aes(x=vine,fill=vine)) + geom_bar() +
  scale_fill_manual(values=cbp[c(7,6)]) + theme_bw() + theme(legend.position = "none") + ggtitle("All Pr
v2 <- ggplot(dryer,aes(x=vine,fill=vine)) + geom_bar() +
  scale_fill_manual(values=cbp[c(7,6)]) + theme_bw() + theme(legend.position = "none") + ggtitle("Dryer
v3 <- ggplot(paci,aes(x=vine,fill=vine)) + geom_bar() +
  scale_fill_manual(values=cbp[c(7,6)]) + theme_bw() + theme(legend.position = "none") + ggtitle("Pacifi
v4 <- ggplot(micro,aes(x=vine,fill=vine)) + geom_bar() +
  scale_fill_manual(values=cbp[c(7,6)]) + theme_bw() + theme(legend.position = "none") + ggtitle("Microw
grid.arrange(v1,v2,v3,v4, ncol=2)
```



For vine reviews, the number of vine reviews is very small. So it isn't likely we can obtain any significant trends from the contents of vine reviews. However, it may still be worthwile to check whether there is any correlation between whether a product has a vine review and how well the product sells/rates.

```
vp1 <- ggplot(all,aes(x=verified_purchase,fill=verified_purchase)) + geom_bar() +
  scale_fill_manual(values=cbp[c(7,6)]) + theme_bw() + theme(legend.position = "none") + ggtitle("All P
vp2 <- ggplot(dryer,aes(x=verified_purchase,fill=verified_purchase)) + geom_bar() +
  scale_fill_manual(values=cbp[c(7,6)]) + theme_bw() + theme(legend.position = "none") + ggtitle("Dryer
vp3 <- ggplot(paci,aes(x=verified_purchase,fill=verified_purchase)) + geom_bar() +
  scale_fill_manual(values=cbp[c(7,6)]) + theme_bw() + theme(legend.position = "none") + ggtitle("Pacif
vp4 <- ggplot(micro,aes(x=verified_purchase,fill=verified_purchase)) + geom_bar() +
  scale_fill_manual(values=cbp[c(7,6)]) + theme_bw() + theme(legend.position = "none") + ggtitle("Micro
grid.arrange(vp1,vp2,vp3,vp4, ncol=2)
```



Since enough verified reviews, just work with verified reviews

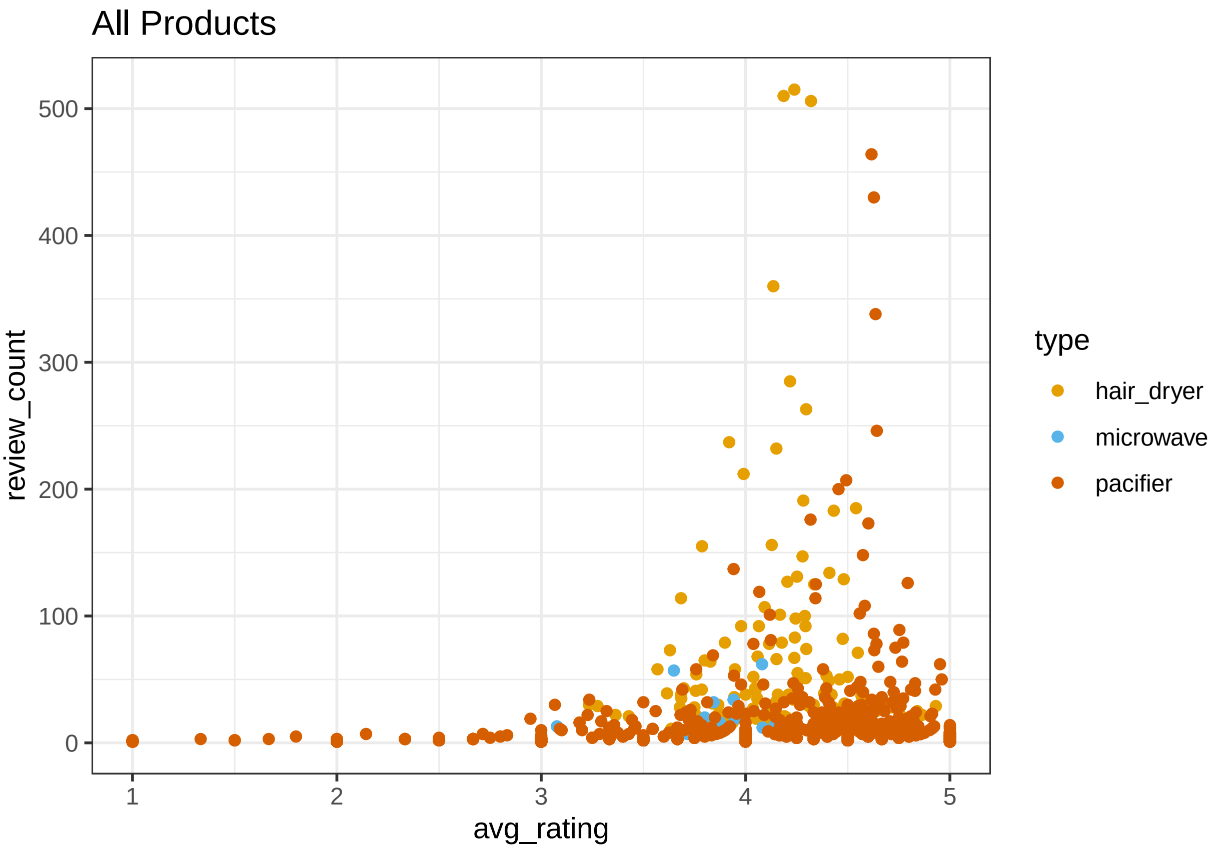
```
all <- read_csv("all_verif.csv")
dryer <- read_csv("hair_dryer_verif.csv")
paci <- read_csv("pacifier_verif.csv")
micro <- read_csv("microwave_verif.csv")
# head(dryer)
# head(paci)
# head(micro)
```

4) Look at relationship of number of reviews and avg star rating for a product_id (there is none)

```
all_products <- all %>% group_by(type,product_id) %>% summarise(avg_rating = mean(star_rating),
                                                                 review_count = n(),
                                                                 helpful_prop = mean(helpful),
                                                                 voted_prop = mean(has_votes))
head(all_products)
```

```
## # A tibble: 6 x 6
## # Groups:   type [1]
##   type       product_id avg_rating review_count helpful_prop voted_prop
##   <chr>      <chr>           <dbl>        <int>        <dbl>      <dbl>
## 1 hair_dryer B000050FDE       3              1           NA       1
## 2 hair_dryer B000052YD1       5             1             0       1
## 3 hair_dryer B00005351F       4             1             1       1
## 4 hair_dryer B00005JGOH       5             1            NA       0
## 5 hair_dryer B00005OOMZ       4.14        360            NA       0.244
## 6 hair_dryer B00006498N       4             1            NA       0
```

```
rs1 <- ggplot(all_products,aes(x=avg_rating,y=review_count, color=type)) + geom_point() +
  scale_color_manual(values=cbp[c(2,3,7)]) + theme_bw() + ggtitle("All Products")
rs1
```



5) Look at change in stats over time

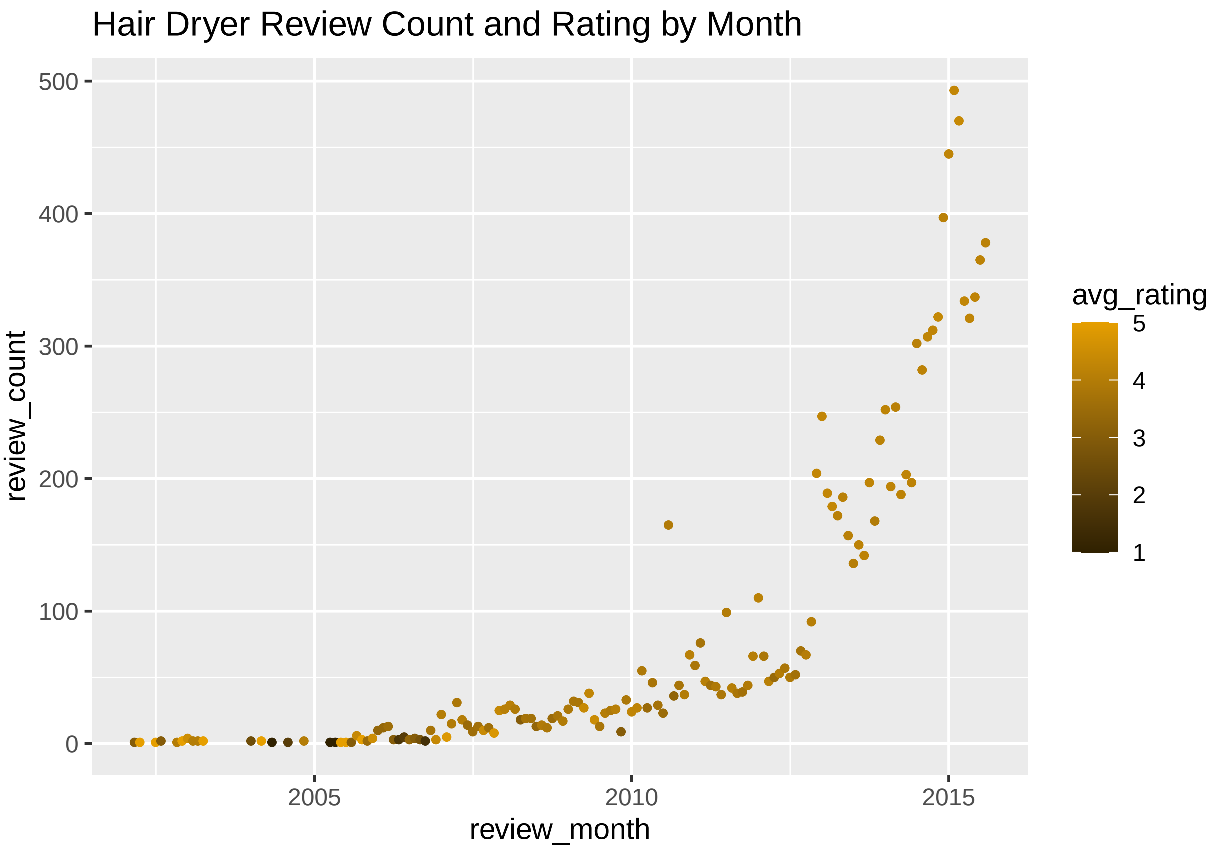
Load in monthly and daily data

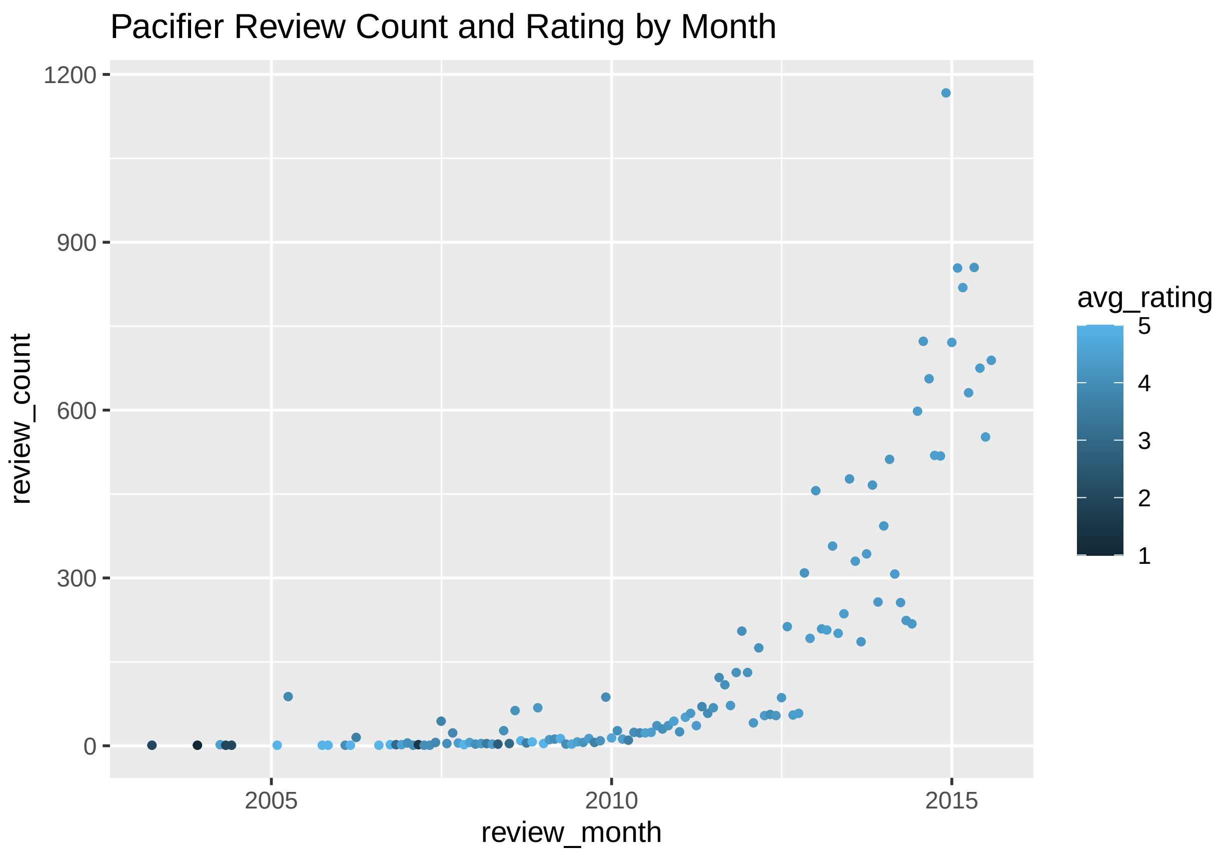
```
dryer_m <- read_csv("monthly_dryer.csv")
paci_m <- read_csv("monthly_paci.csv")
micro_m <- read_csv("monthly_micro.csv")
dryer_d <- read_csv("daily_dryer.csv")
paci_d <- read_csv("daily_paci.csv")
micro_d <- read_csv("daily_micro.csv")
head(dryer_m)
```

```
## # A tibble: 6 x 11
##   review_month review_count product_titles review_headlines review_bodies
##   <date>              <dbl> <chr>          <chr>            <chr>
## 1 2002-03-01              1 conair corp p~ some pluses, so~ this is my o~
## 2 2002-04-01              1 conair corp p~ excellent for f~ this hairdry~
## 3 2002-07-01              1 conair corp p~ the best dryer!~ i love this ~
## 4 2002-08-01              2 revlon 1875w ~ great hair! | d~ i just purch~
## 5 2002-11-01              1 revlon 1875w ~ close to perfect overall, i a~
## 6 2002-12-01              2 conair corp p~ good buy | exce~ one of the m~
## # ... with 6 more variables: star_ratings <chr>, avg_rating <dbl>,
## #   impact_pos <dbl>, impact_neg <dbl>, impact_overall <dbl>, cum_rating <dbl>
```

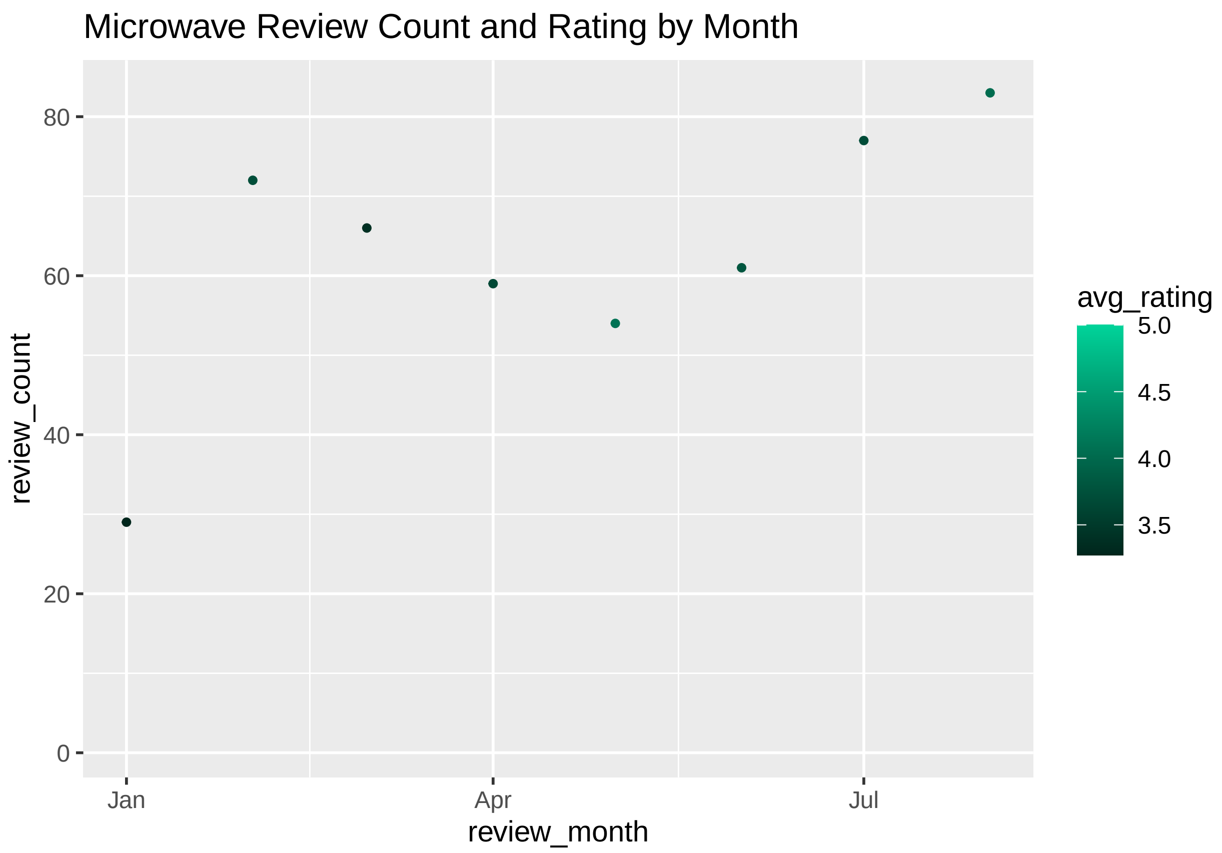
5a) Looking at all products together not that insightful

```
dc1 <- ggplot(dryer_m,aes(x=review_month,y=review_count,color=avg_rating)) + geom_point(size=1) +
  scale_color_gradient(low = "#302100",high = "#E69F00") + ggtitle("Hair Dryer Review Count and Rating b
dc2 <- ggplot(paci_m,aes(x=review_month,y=review_count,color=avg_rating)) + geom_point(size=1) +
  scale_color_gradient(low = "#122733",high = "#56B4E9") + ggtitle("Pacifier Review Count and Rating by
dc3 <- ggplot(micro_m,aes(x=review_month,y=review_count,color=avg_rating)) + geom_point(size=1) +
  scale_color_gradient(low = "#00261c",high = "#00d49a") + ggtitle("Microwave Review Count and Rating by
dc1; dc2; dc3
```





```
## Warning: Removed 1 rows containing missing values (geom_point).
```



Look at data for most product_ids with most reviews

```
dryer_review_ct <- dryer %>% group_by(product_id) %>% summarise(review_count=n()) %>% arrange(desc(revi
head(dryer_review_ct,30) # filter for at least 100 reviews
```

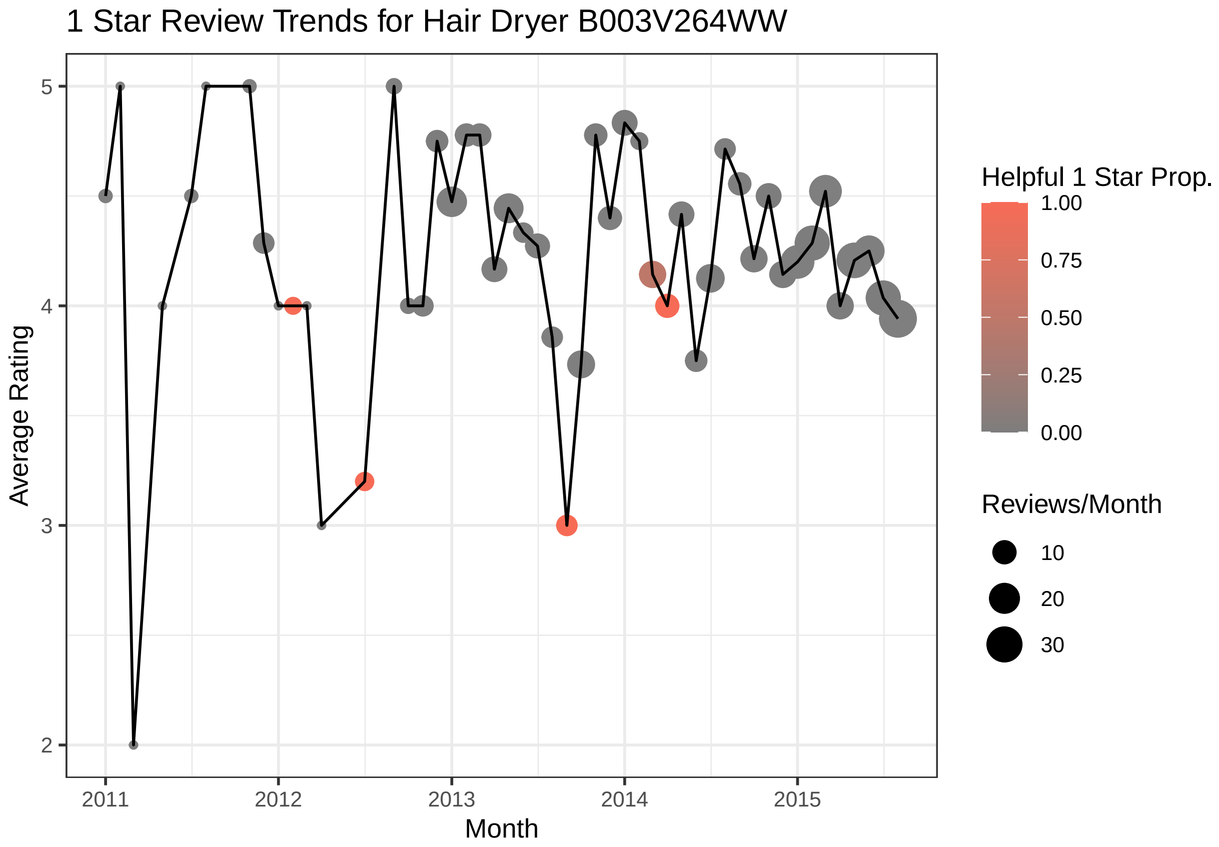
```
## # A tibble: 30 x 2
##    product_id review_count
##    <chr>             <int>
##  1 B003V264WW          515
##  2 B0009XH6TG          510
##  3 B00132ZG3U          506
##  4 B00005OOMZ          360
##  5 B000A3I2X4          285
##  6 B000R80ZTQ          263
##  7 B001UE7D2I          237
##  8 B001QTW2FK          232
##  9 B0009XH6WI          212
## 10 B0009XH6V4          191
## # ... with 20 more rows
```

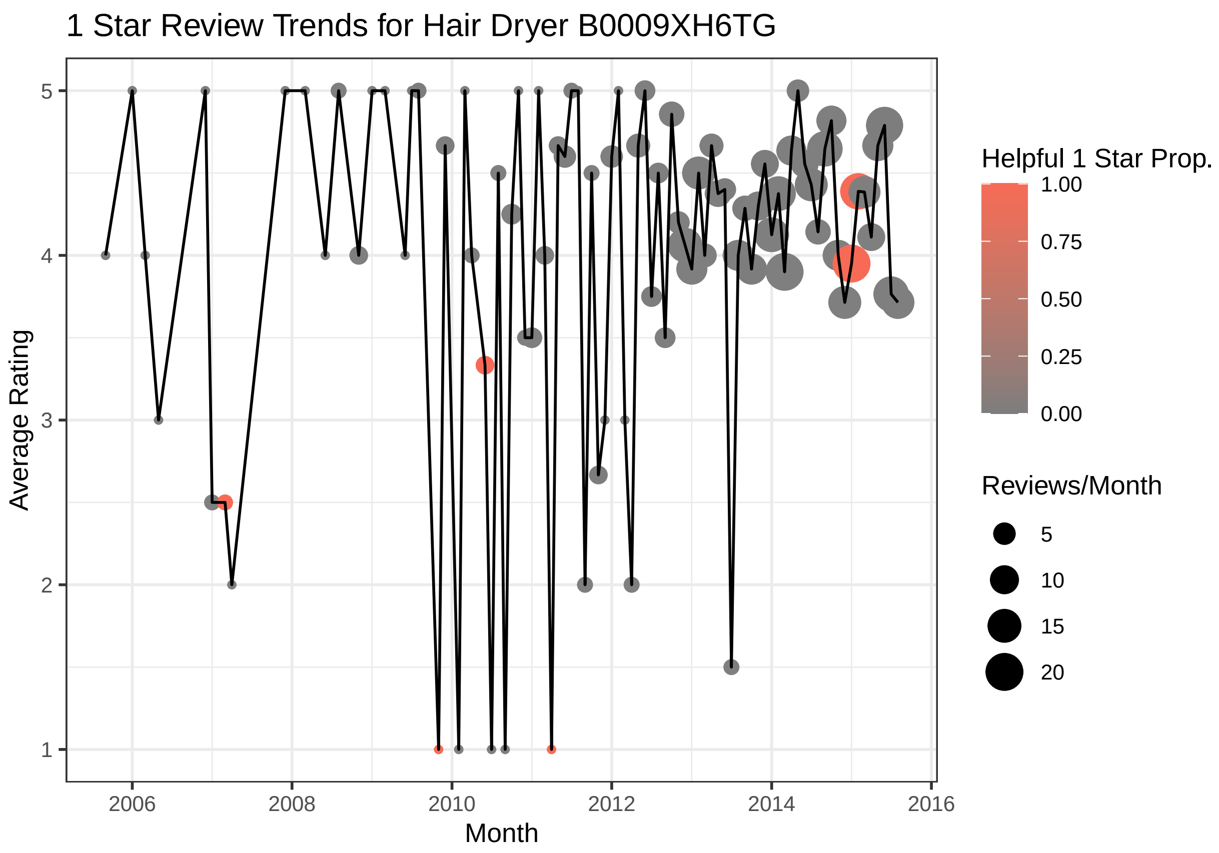
```
pop_dryer <- dryer %>% group_by(product_id) %>% mutate(review_count=n()) %>% arrange(desc(review_count)
pop_micro <- micro %>% group_by(product_id) %>% mutate(review_count=n()) %>% arrange(desc(review_count)
pop_paci <- paci %>% group_by(product_id) %>% mutate(review_count=n()) %>% arrange(desc(review_count))
# pop_dryer
```

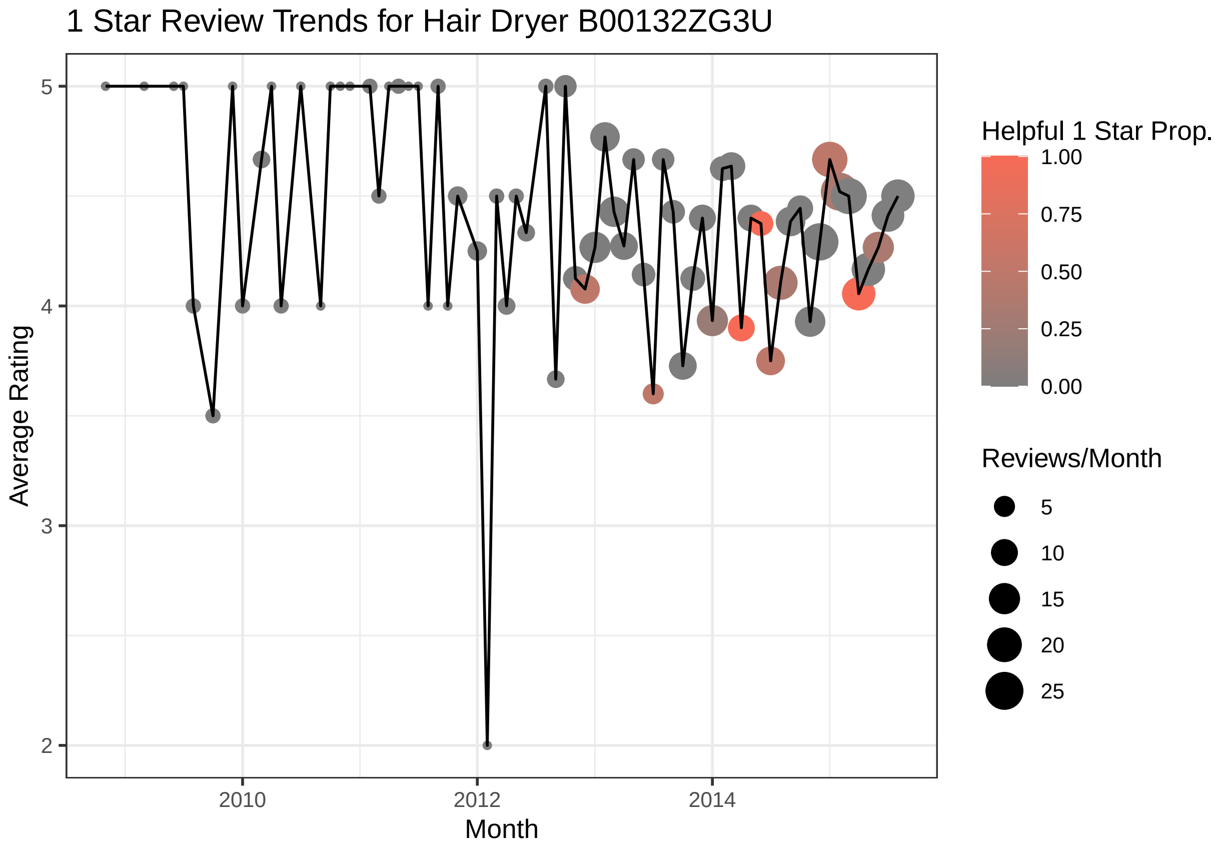
5b) look at helpful 1 and 5 star

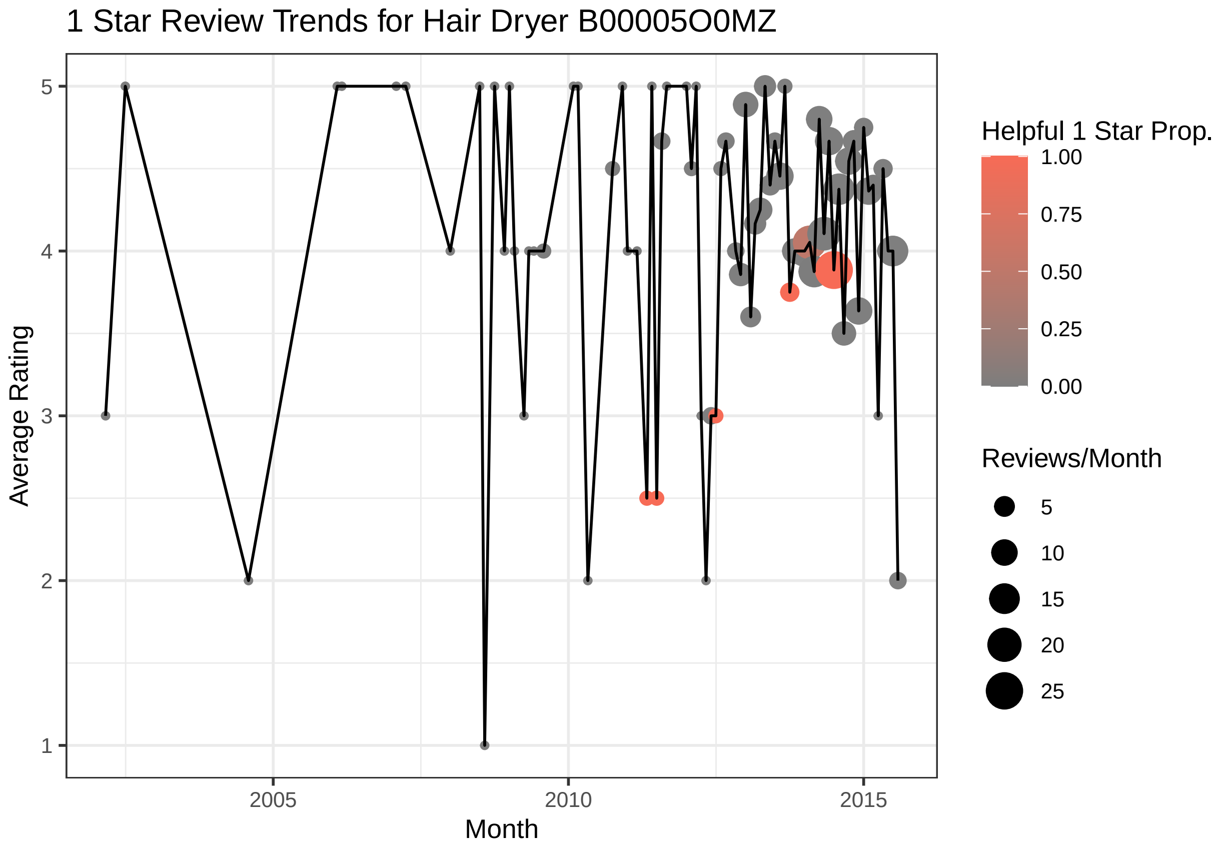
however these do seem to occur at dips in avg rating as expected

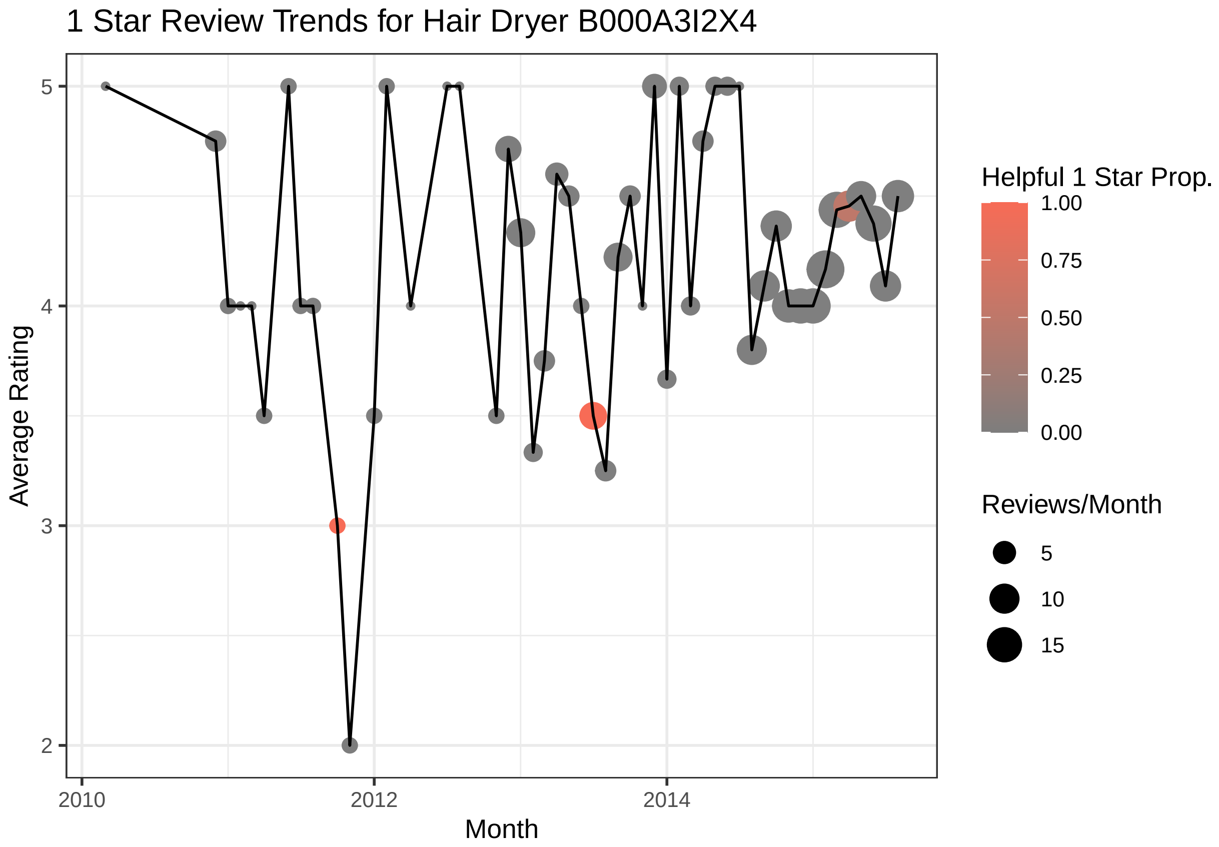
```r
for(product in unique(pop_dryer$product_id)[1:5]){
  prod_df <- pop_dryer[pop_dryer$product_id == product,] %>% group_by(review_month,product_id) %>%
    summarise(review_count=n(), avg_rating=mean(star_rating),
              impact_1_prop=mean(impact_star == "helpful 1 star",na.rm=T),
              impact_5_prop=mean(impact_star == "helpful 5 star",na.rm=T))
  g <- ggplot(data=prod_df, aes(x=review_month,y=avg_rating)) + geom_point(aes(size=review_count,color=
    scale_color_gradient(low = "#7d7d7d",high = "#f76b56",name = "Helpful 1 Star Prop.") + labs(size="R
    ggtitle(paste("1 Star Review Trends for Hair Dryer",prod_df$product_id)) +
    theme_bw() + theme(text = element_text(size=10))
  plot(g)
}
```




1 Star Review Trends for Hair Dryer B0009XH6TG
Average Rating
5
4
3
2
1
2006
2008
2010
2012
2014
2016
Month
Helpful 1 Star Prop.
1.00
0.75
0.50
0.25
0.00
Reviews/Month
5
10
15
20

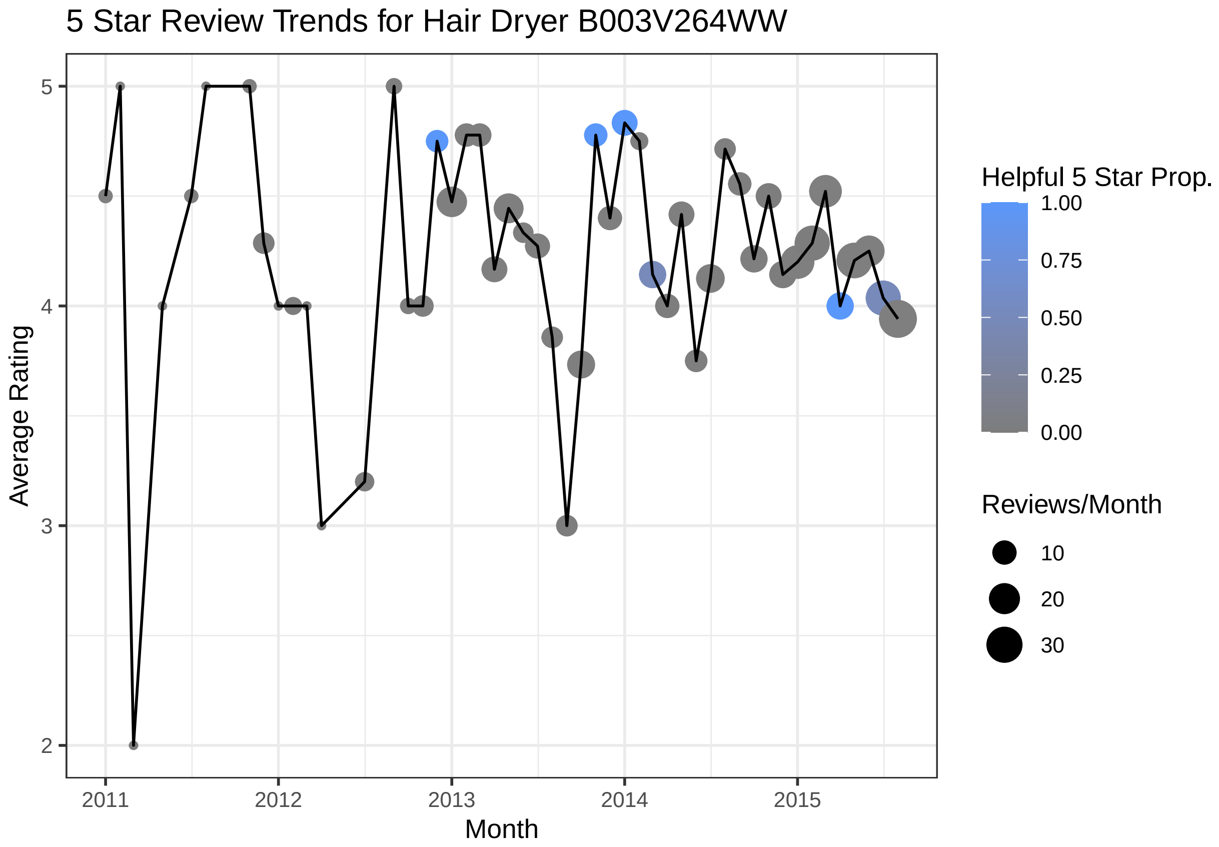

1 Star Review Trends for Hair Dryer B00132ZG3U
Average Rating
5
4
3
2
Month
2010
2012
2014
Helpful 1 Star Prop.
1.00
0.75
0.50
0.25
0.00
Reviews/Month
5
10
15
20
25

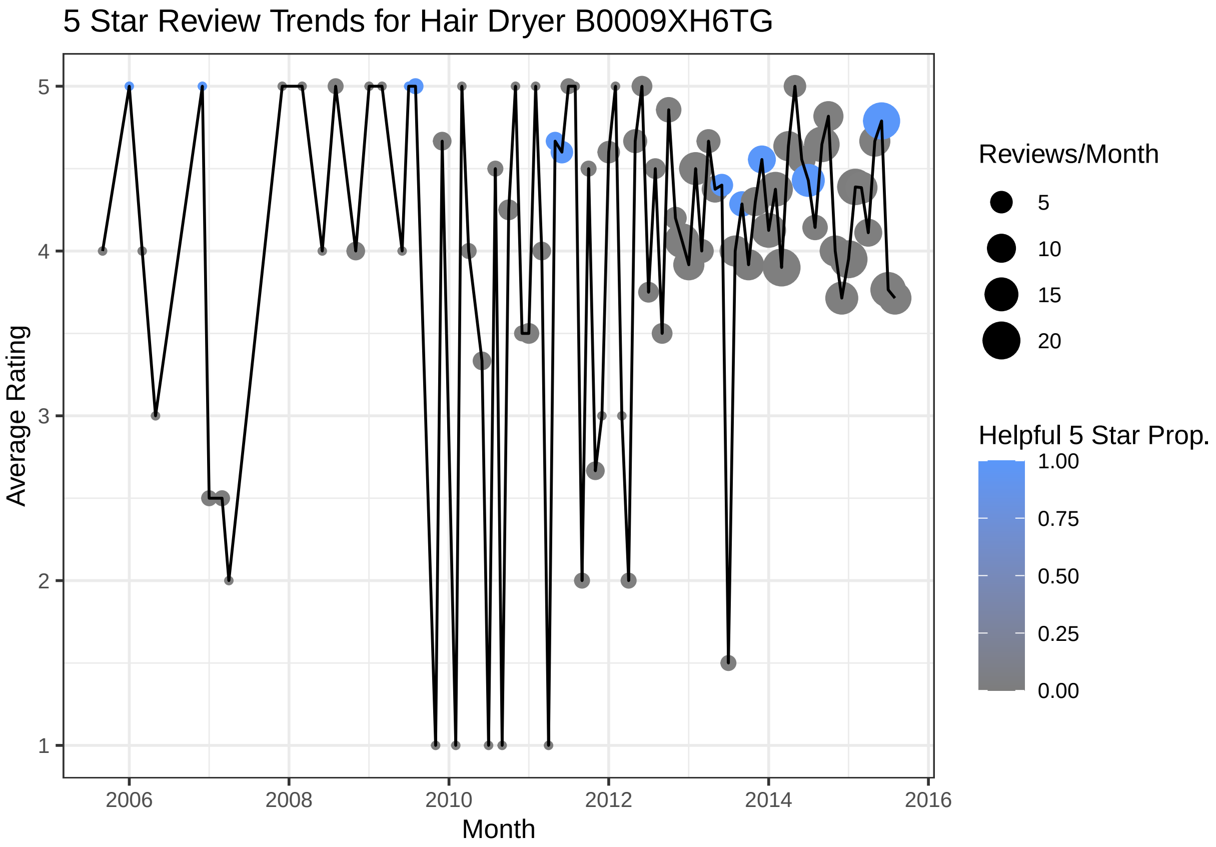

1 Star Review Trends for Hair Dryer B00005O0MZ
Average Rating
5
4
3
2
1
Month
2005
2010
2015
Helpful 1 Star Prop.
1.00
0.75
0.50
0.25
0.00
Reviews/Month
5
10
15
20
25

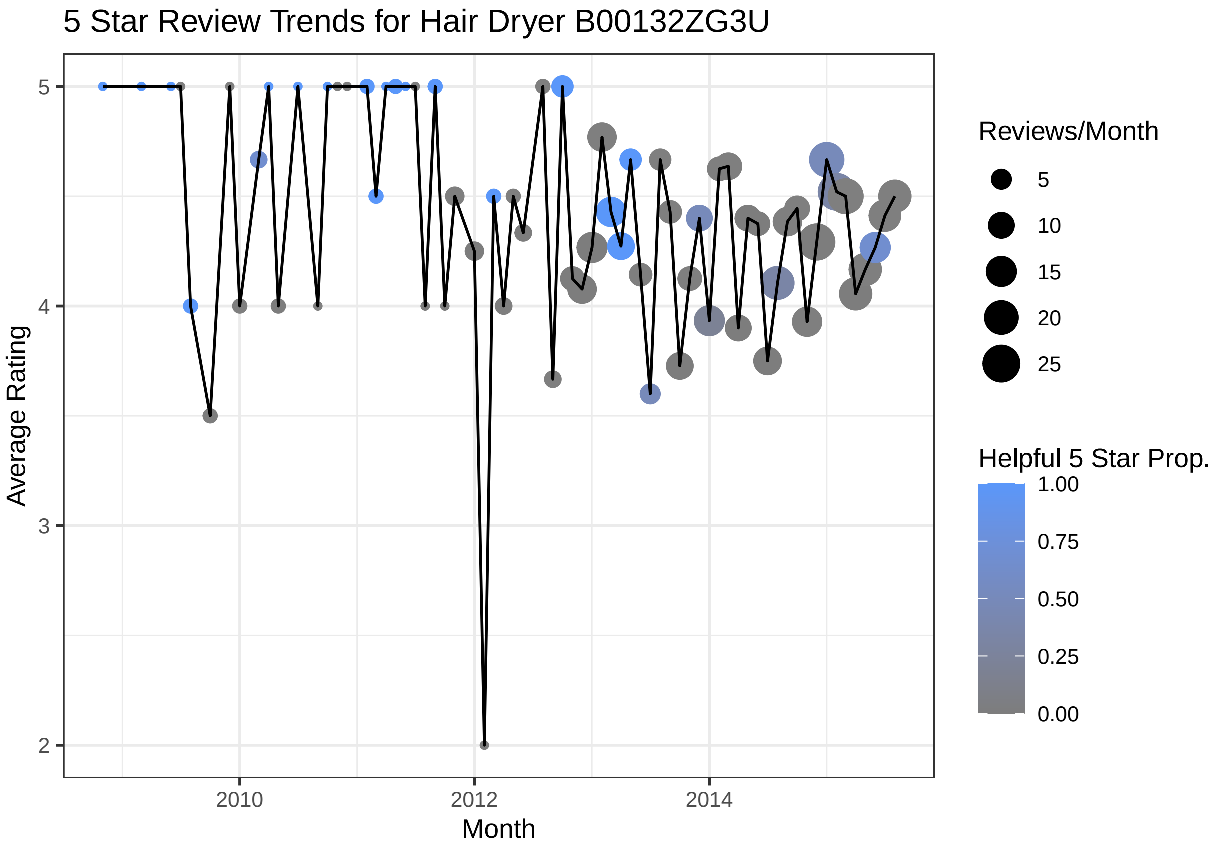


Repeating for 5 star reviews, we also see that impactful reviews appear near the “peaks”

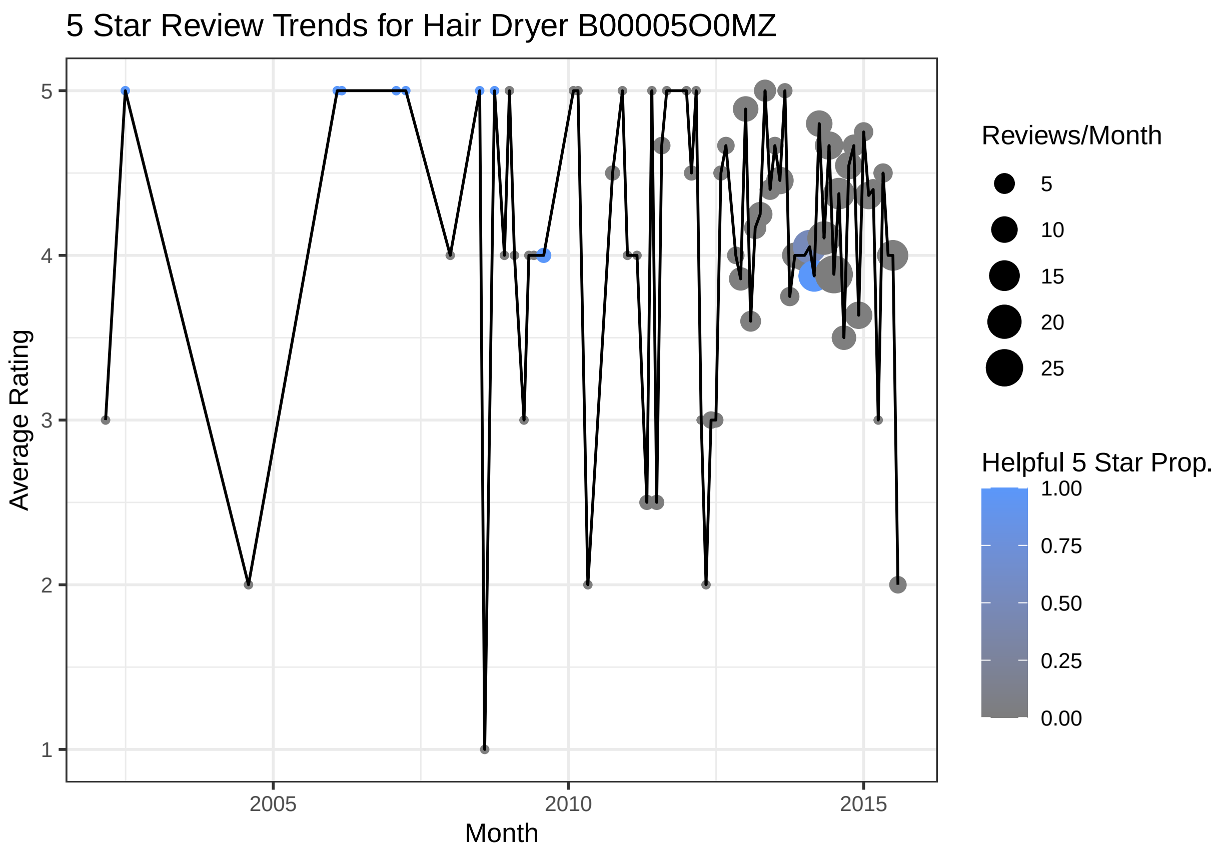
```
for(product in unique(pop_dryer$product_id)[1:5]){
  prod_df <- pop_dryer[pop_dryer$product_id == product,] %>% group_by(review_month,product_id) %>%
    summarise(review_count=n(), avg_rating=mean(star_rating),
              impact_1_prop=mean(impact_star == "helpful 1 star",na.rm=T),
              impact_5_prop=mean(impact_star == "helpful 5 star",na.rm=T))
  g <- ggplot(data=prod_df, aes(x=review_month,y=avg_rating)) + geom_point(aes(size=review_count,color=
    scale_color_gradient(low = "#7d7d7d",high = "#5a97fa",name = "Helpful 5 Star Prop.") + labs(size="R
    ggtitle(paste("5 Star Review Trends for Hair Dryer",prod_df$product_id)) +
    theme_bw() + theme(text = element_text(size=10))
  plot(g)
}
```

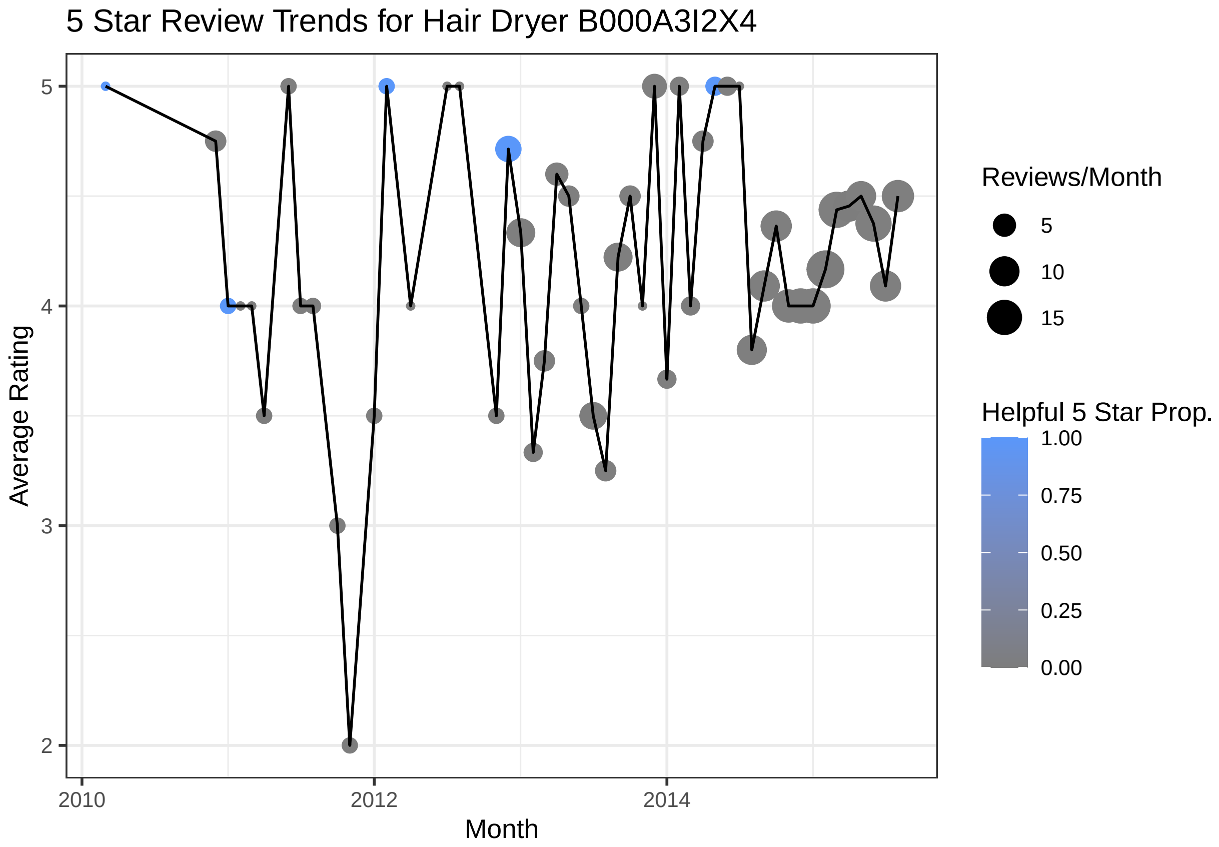

5 Star Review Trends for Hair Dryer B003V264WW
Average Rating
5
4
3
2
Month
2011
2012
2013
2014
2015
Helpful 5 Star Prop.
1.00
0.75
0.50
0.25
0.00
Reviews/Month
10
20
30


5 Star Review Trends for Hair Dryer B0009XH6TG
Average Rating
5
4
3
2
1
Month
2006
2008
2010
2012
2014
2016
Reviews/Month
5
10
15
20
Helpful 5 Star Prop.
1.00
0.75
0.50
0.25
0.00



5 Star Review Trends for Hair Dryer B00132ZG3U
Average Rating
5
4
3
2
2010
2012
2014
Month
Reviews/Month
5
10
15
20
25
Helpful 5 Star Prop.
1.00
0.75
0.50
0.25
0.00

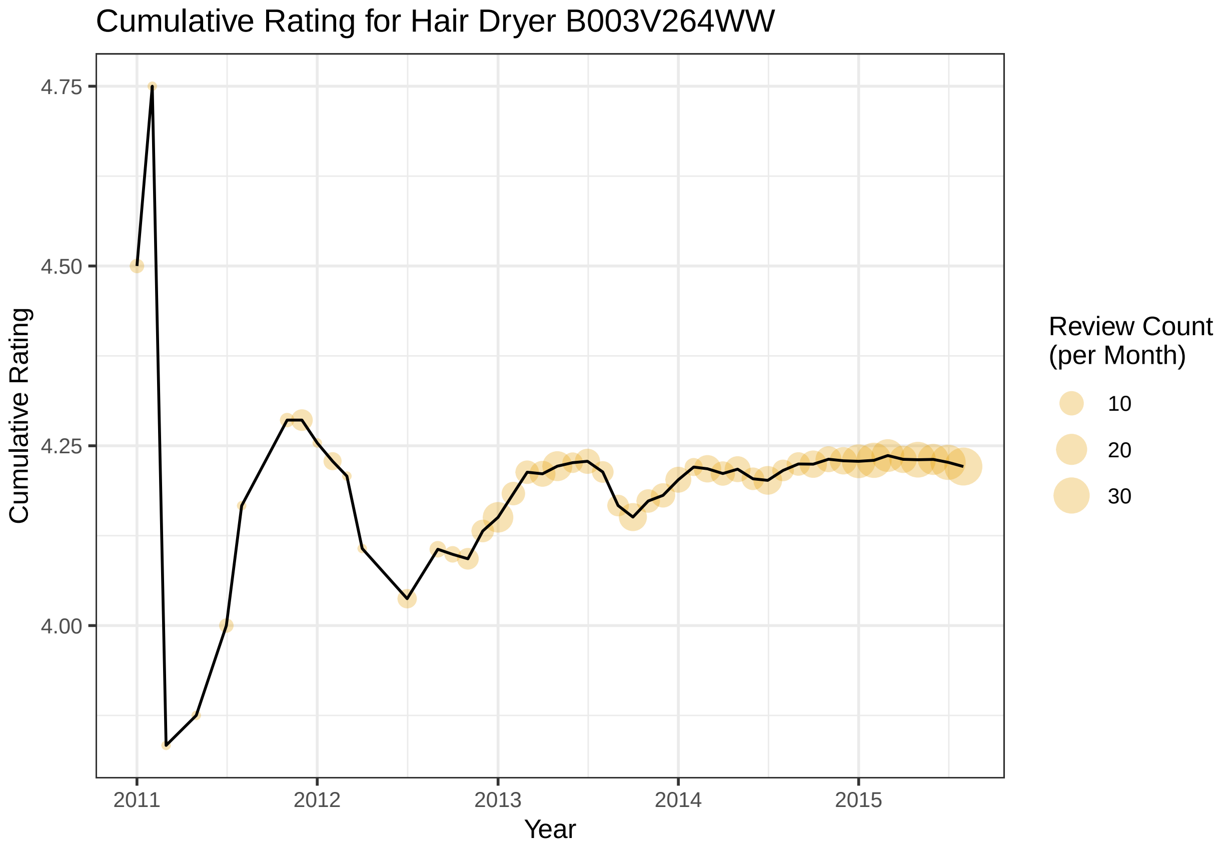

5 Star Review Trends for Hair Dryer B00005O0MZ
Average Rating
1
2
3
4
5
Month
2005
2010
2015
Reviews/Month
5
10
15
20
25
Helpful 5 Star Prop.
1.00
0.75
0.50
0.25
0.00

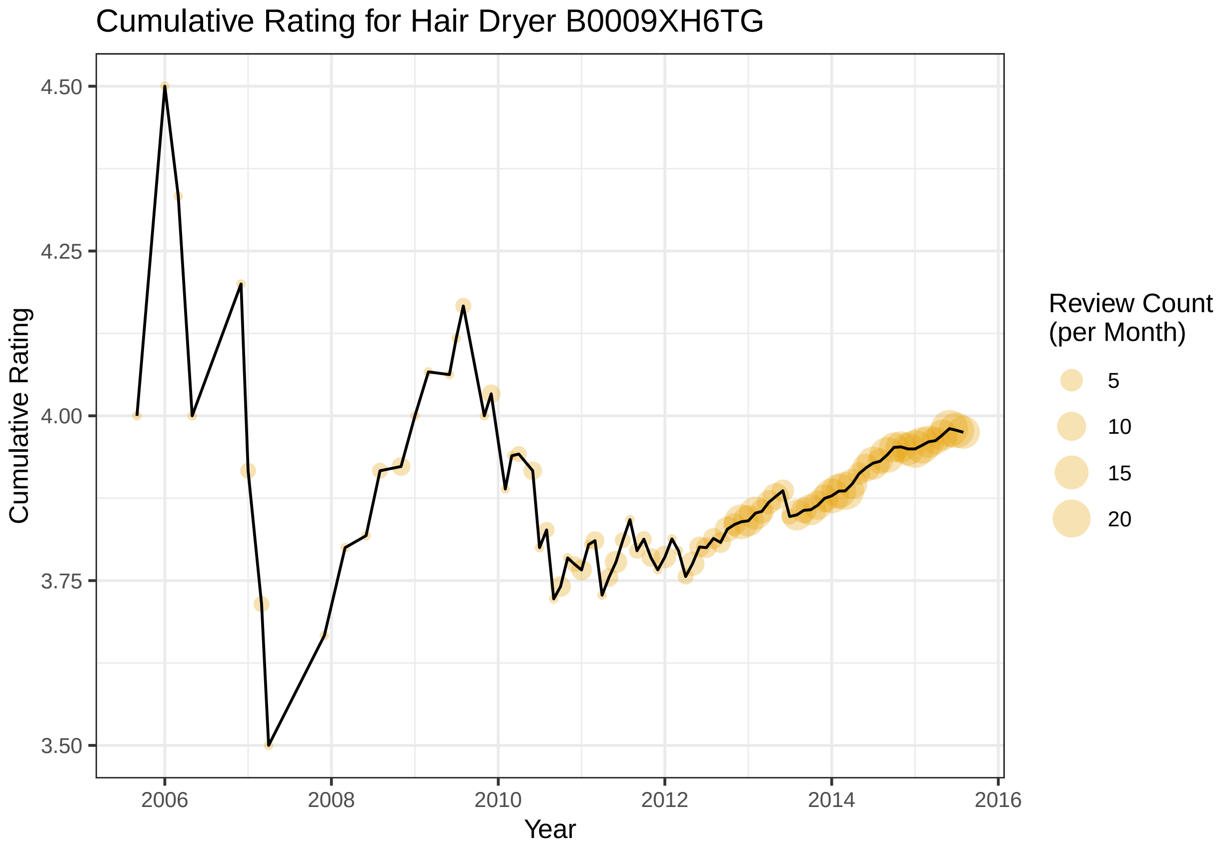


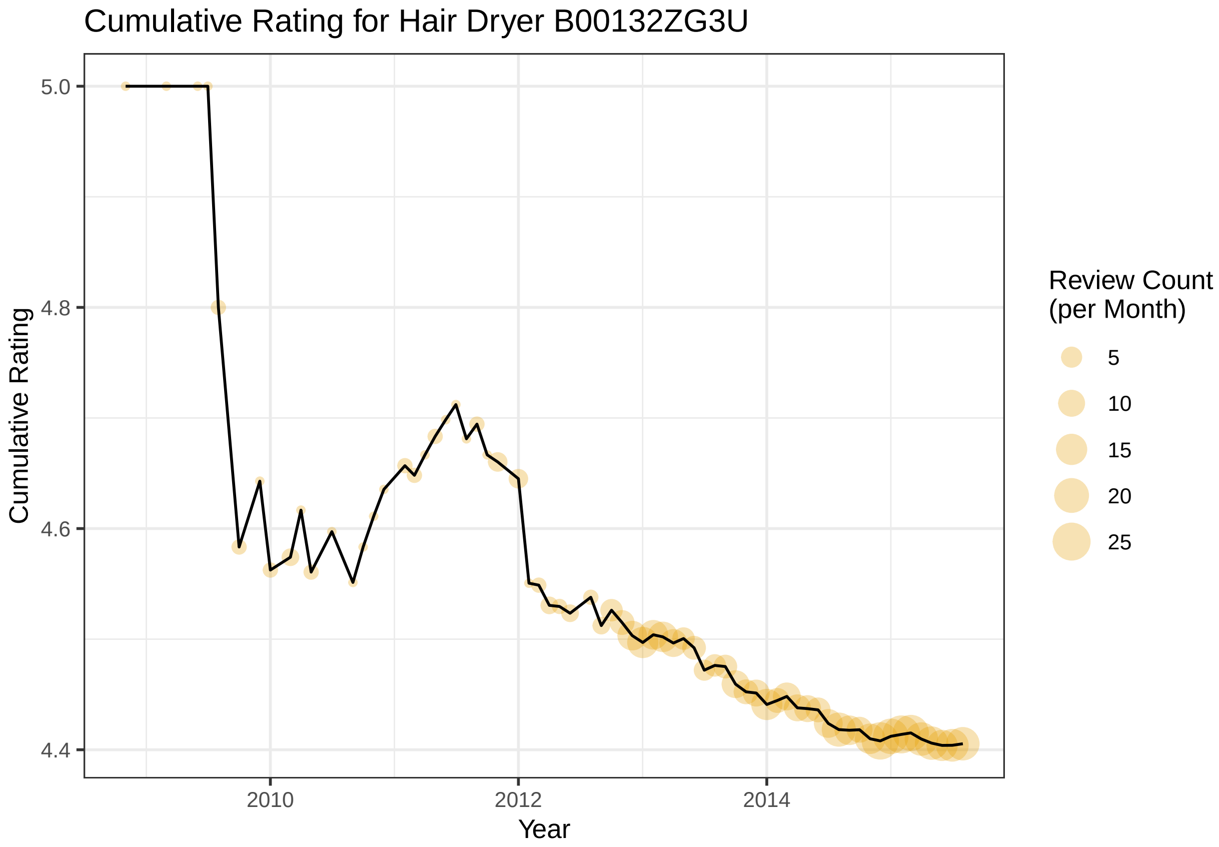
5c) look at review_count vs cumulative rating

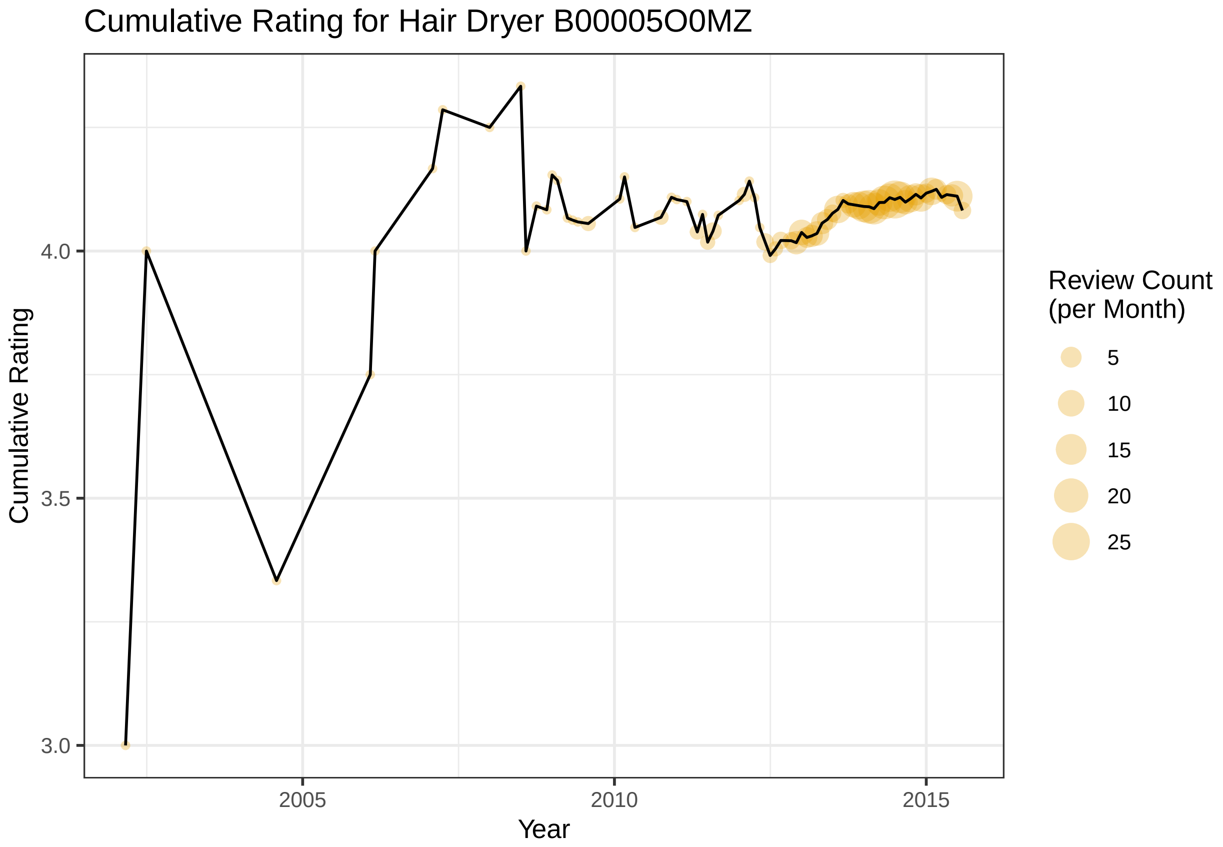
```
cum_rtg_plots <- function(df,col,num,type){
  for(product in unique(df$product_id)[1:num]){
    prod_df <- df[df$product_id == product,] %>% group_by(review_month,product_id) %>%
      summarise(review_count=n(),avg_star = mean(star_rating)) %>% ungroup() %>%  mutate(cum_star = cumm
    g <- ggplot(data=prod_df, aes(x=review_month,y=cum_star)) + geom_point(aes(size=review_count),alpha
      ggtitle(paste("Cumulative Rating for",type,prod_df$product_id)) + ylab("Cumulative Rating") + xla
      theme_bw() + theme(text = element_text(size=10))
    plot(g)
  }
}
```

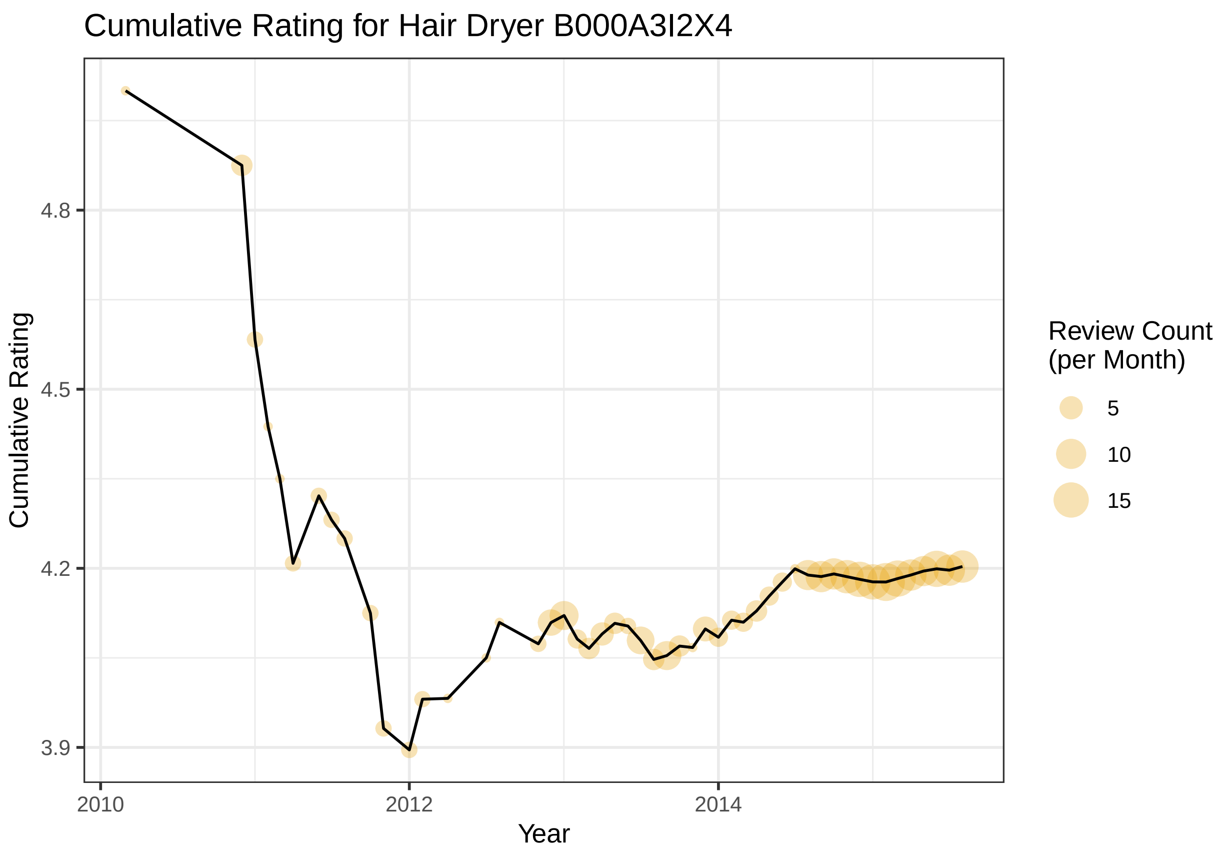
```
cum_rtg_plots(pop_dryer,"#E69F00",5,"Hair Dryer")
```


Cumulative Rating for Hair Dryer B003V264WW
Cumulative Rating
4.75
4.50
4.25
4.00
Year
2011
2012
2013
2014
2015
Review Count
(per Month)
10
20
30

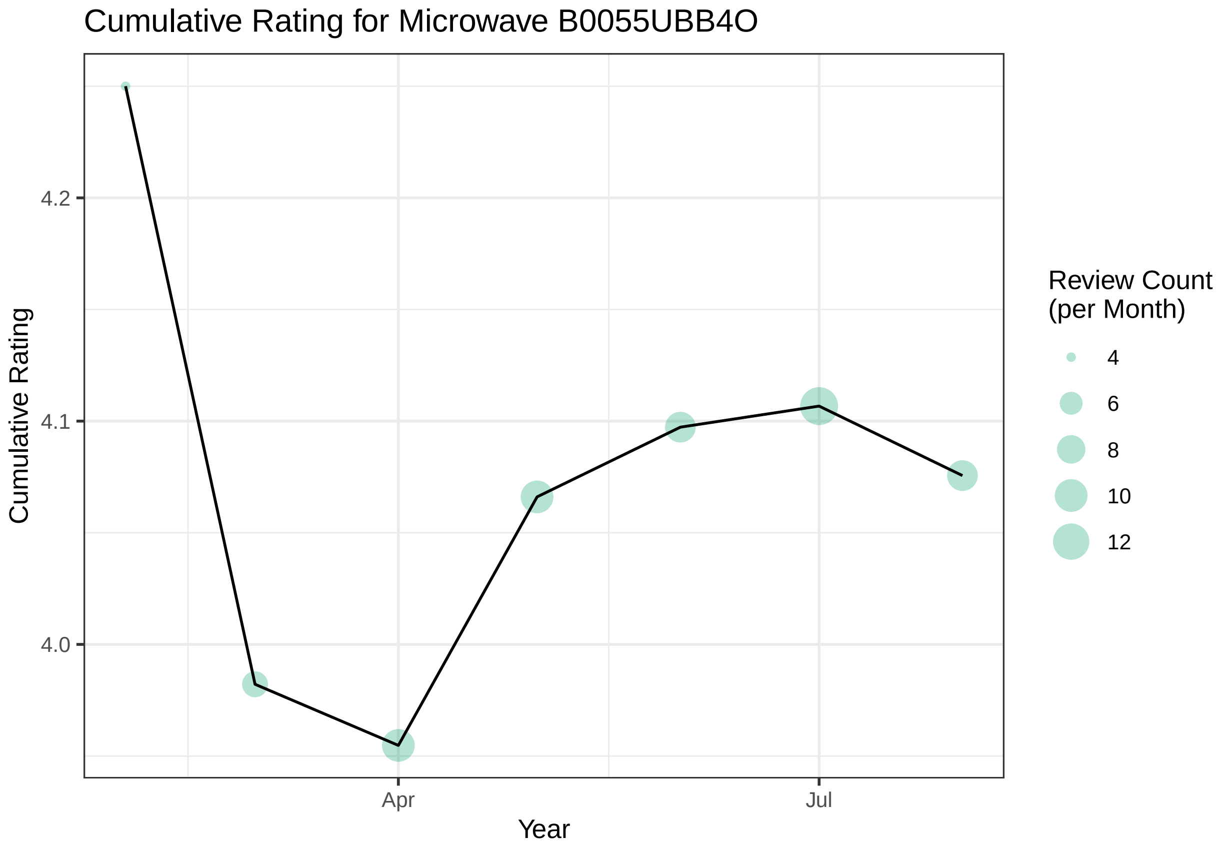

Cumulative Rating for Hair Dryer B0009XH6TG
Cumulative Rating
4.50
4.25
4.00
3.75
3.50
2006
2008
2010
2012
2014
2016
Year
Review Count
(per Month)
5
10
15
20

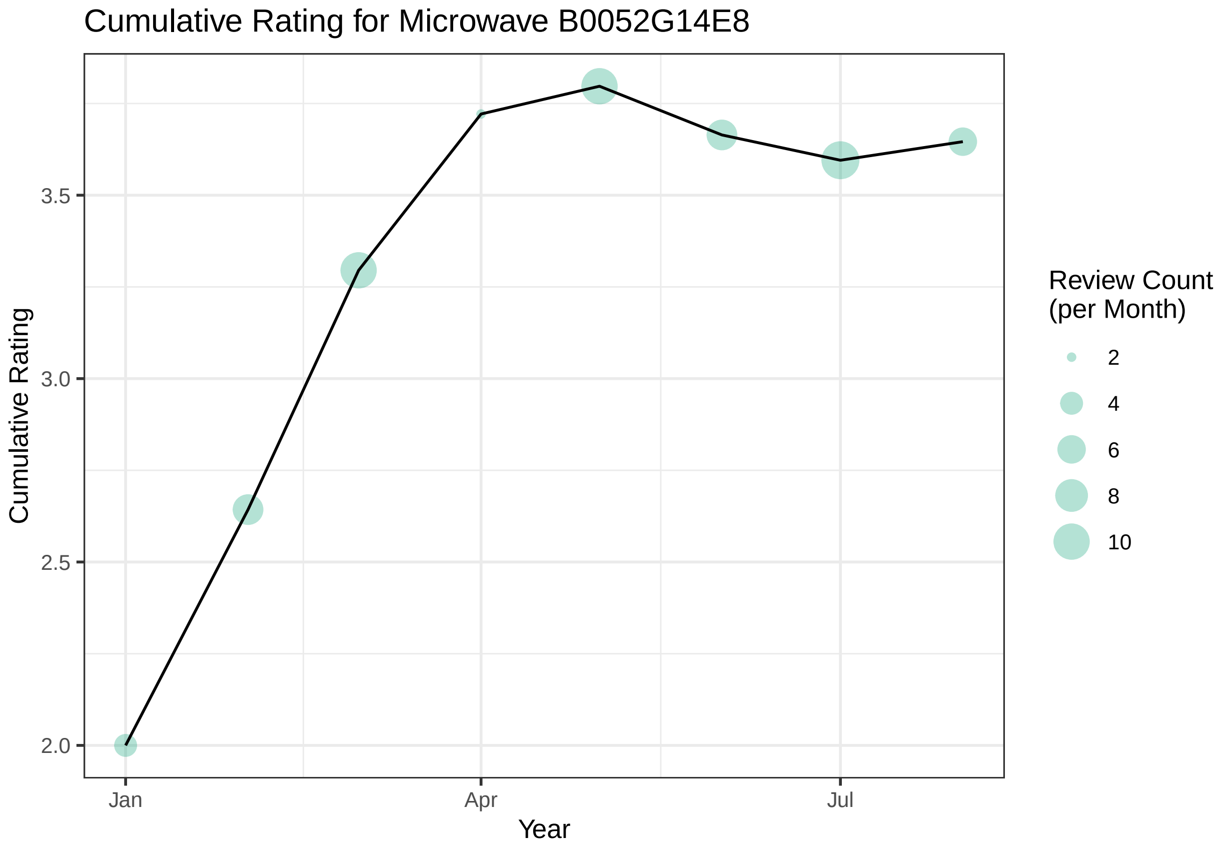

Cumulative Rating for Hair Dryer B00132ZG3U
Cumulative Rating
5.0
4.8
4.6
4.4
2010
2012
2014
Year
Review Count
(per Month)
5
10
15
20
25

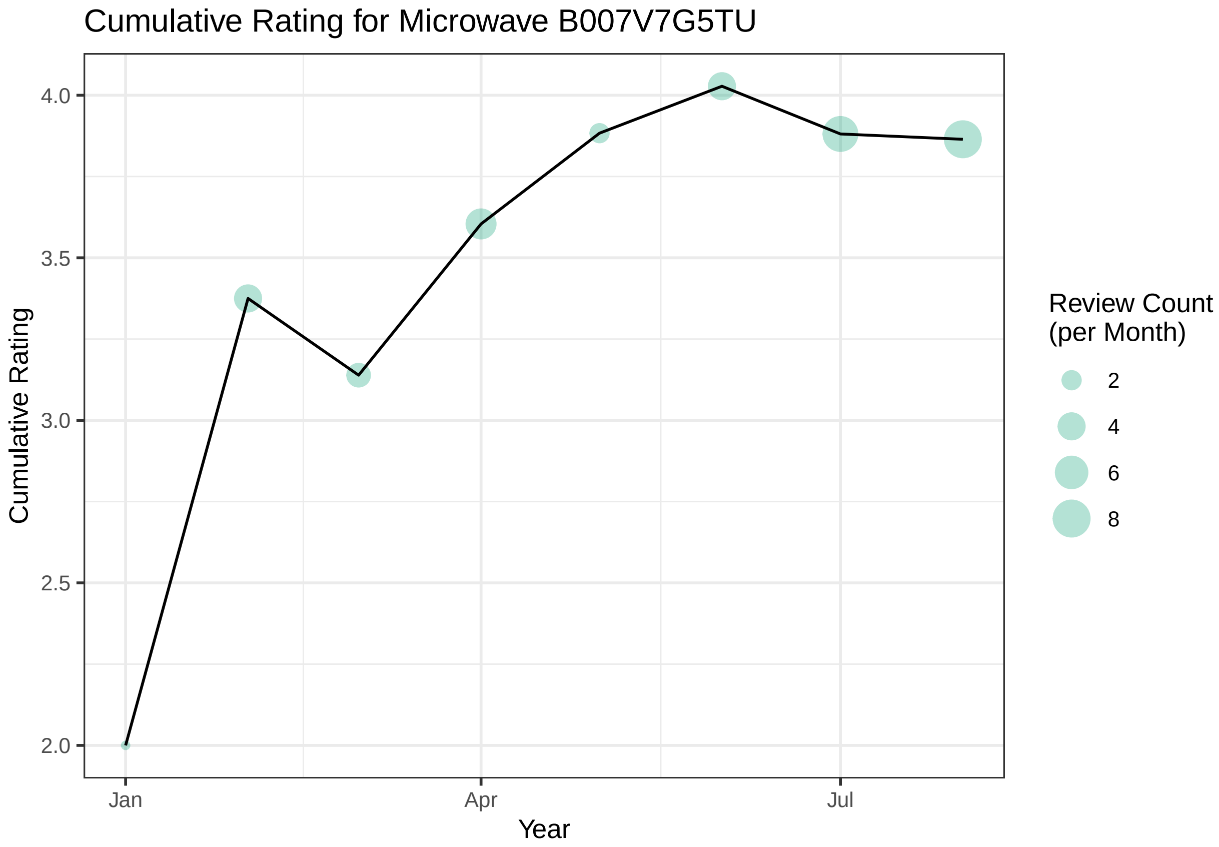

Cumulative Rating for Hair Dryer B00005O0MZ
Cumulative Rating
4.0
3.5
3.0
2005
2010
2015
Year
Review Count
(per Month)
5
10
15
20
25

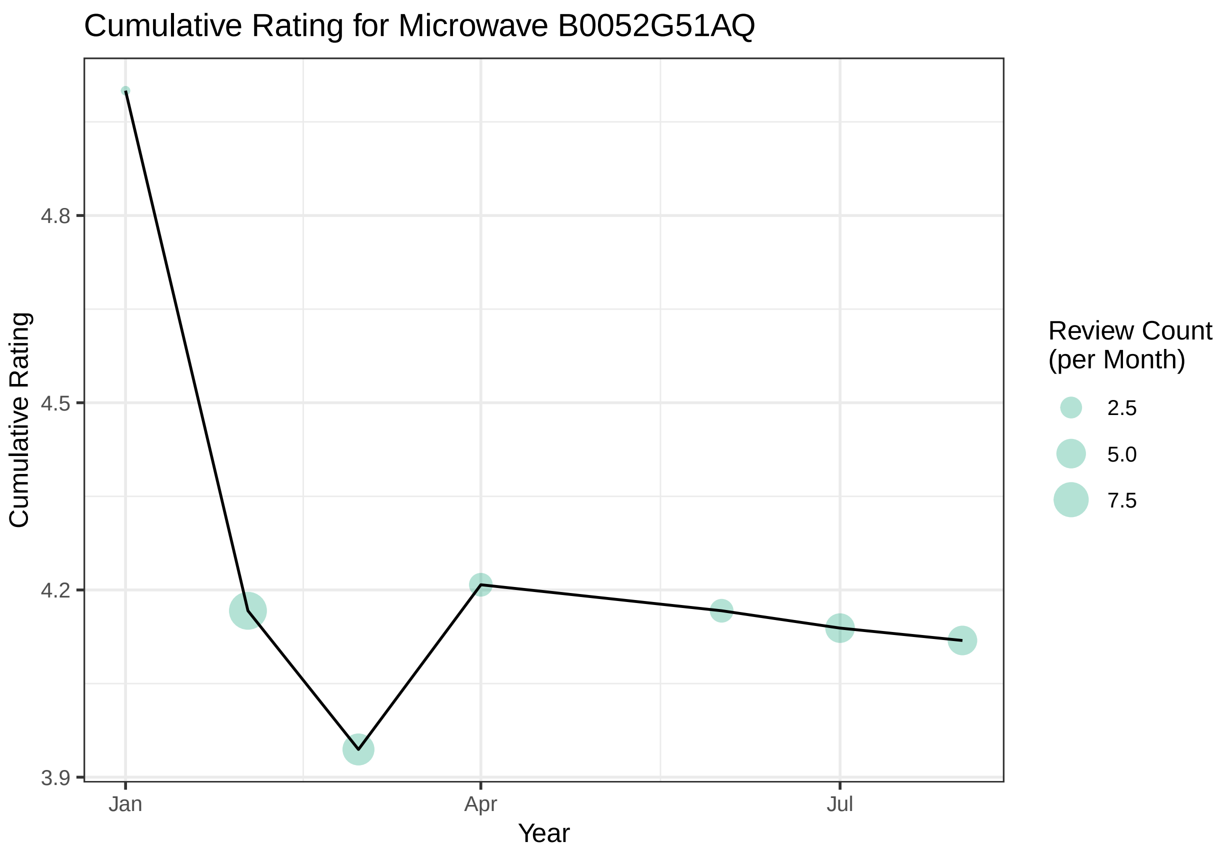


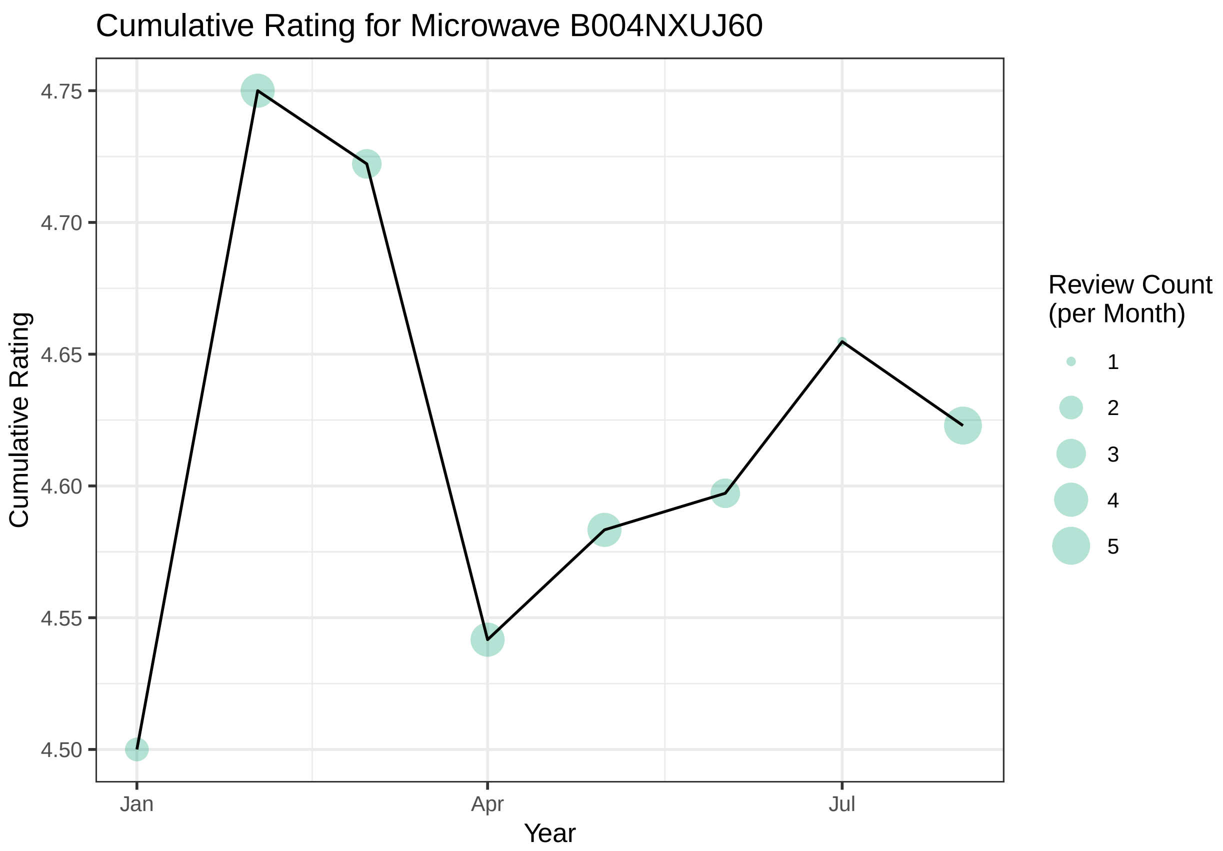
```
cum_rtg_plots(pop_micro,"#009E73",5,"Microwave")
```


Cumulative Rating for Microwave B0055UBB4O
Cumulative Rating
4.2
4.1
4.0
Apr
Jul
Year
Review Count
(per Month)
4
6
8
10
12

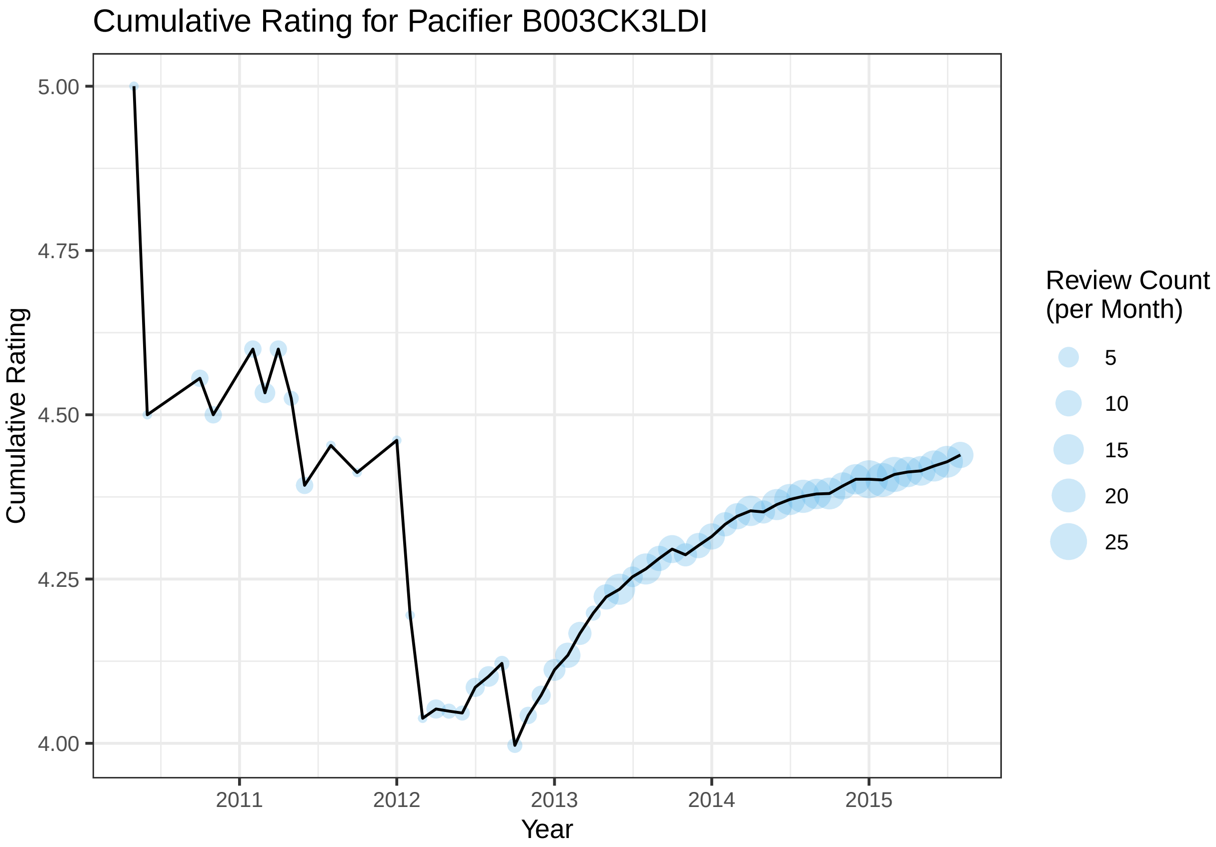

Cumulative Rating for Microwave B0052G14E8
Cumulative Rating
3.5
3.0
2.5
2.0
Jan
Apr
Jul
Year
Review Count
(per Month)
2
4
6
8
10

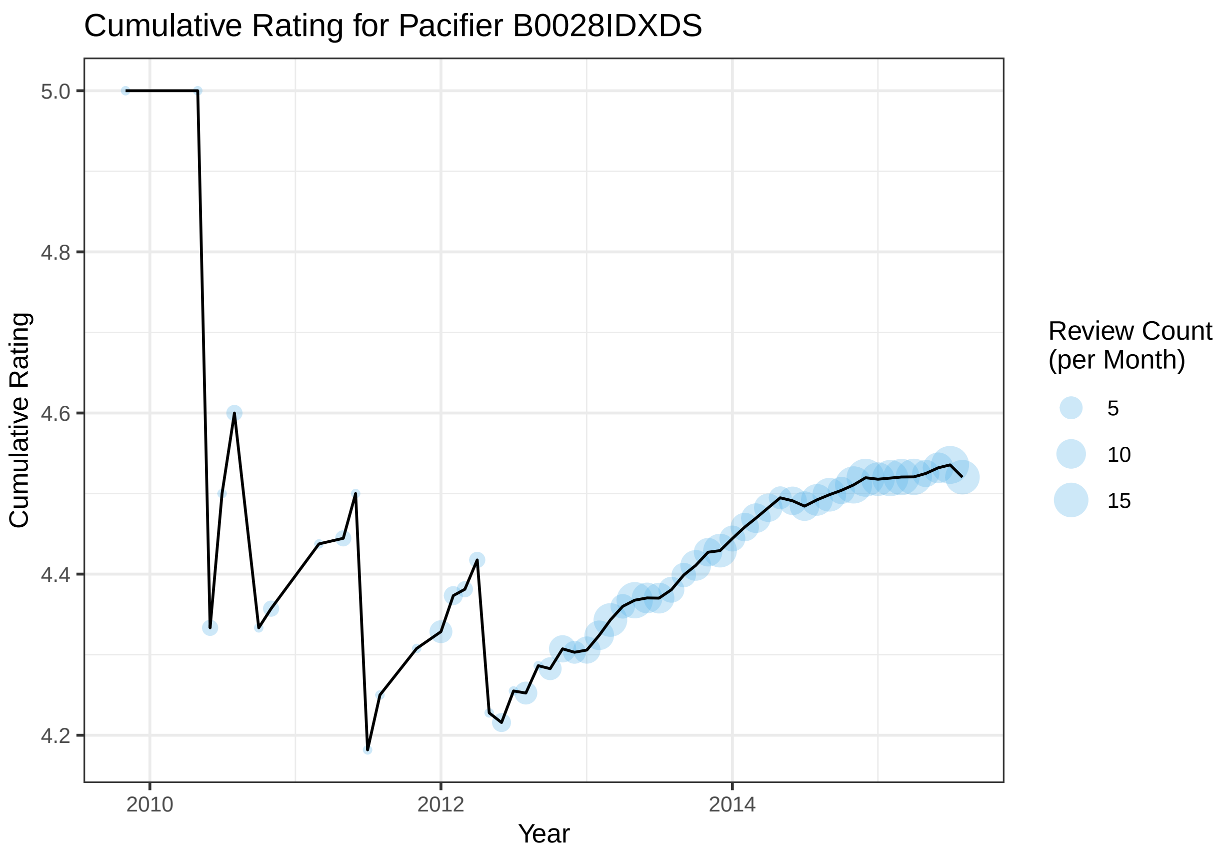

Cumulative Rating for Microwave B007V7G5TU
Cumulative Rating
4.0
3.5
3.0
2.5
2.0
Jan
Apr
Jul
Year
Review Count
(per Month)
2
4
6
8

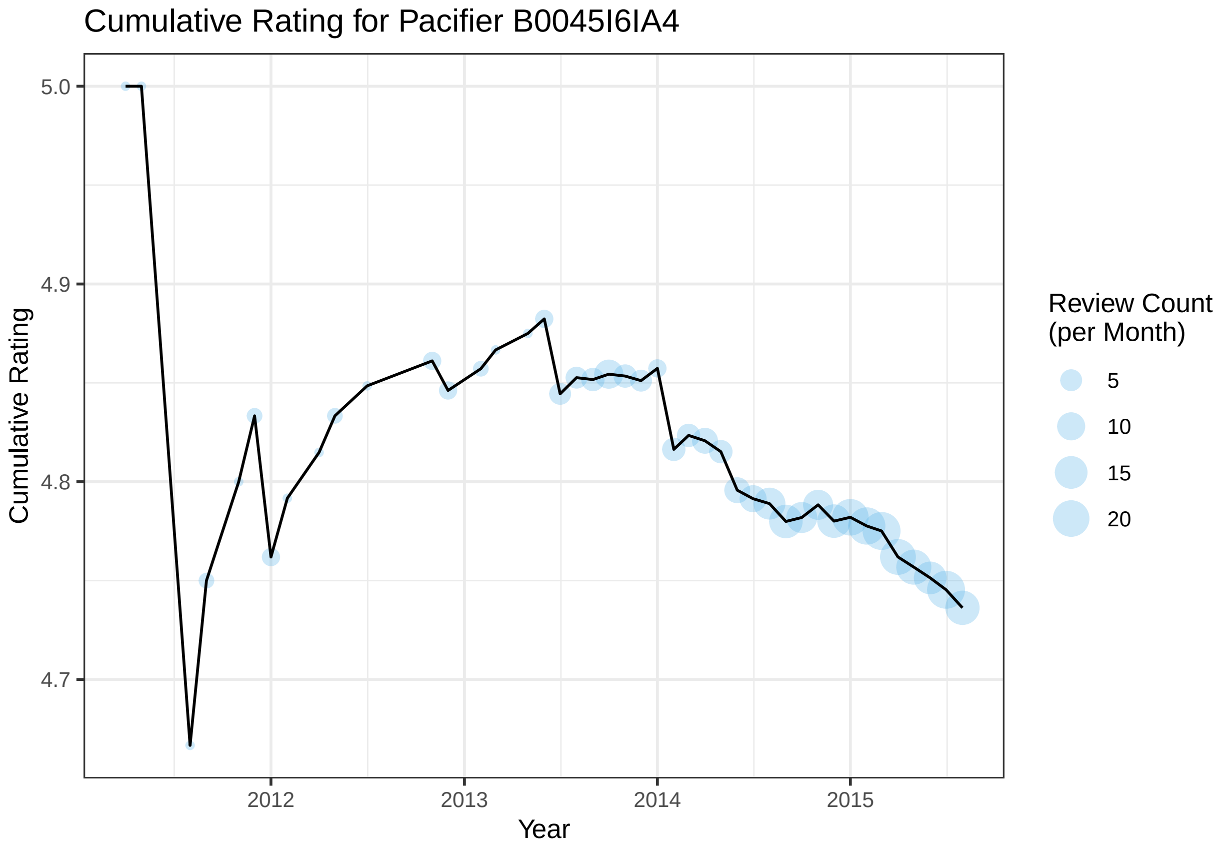

Cumulative Rating for Microwave B0052G51AQ
Cumulative Rating
4.8
4.5
4.2
3.9
Jan
Apr
Jul
Year
Review Count
(per Month)
2.5
5.0
7.5

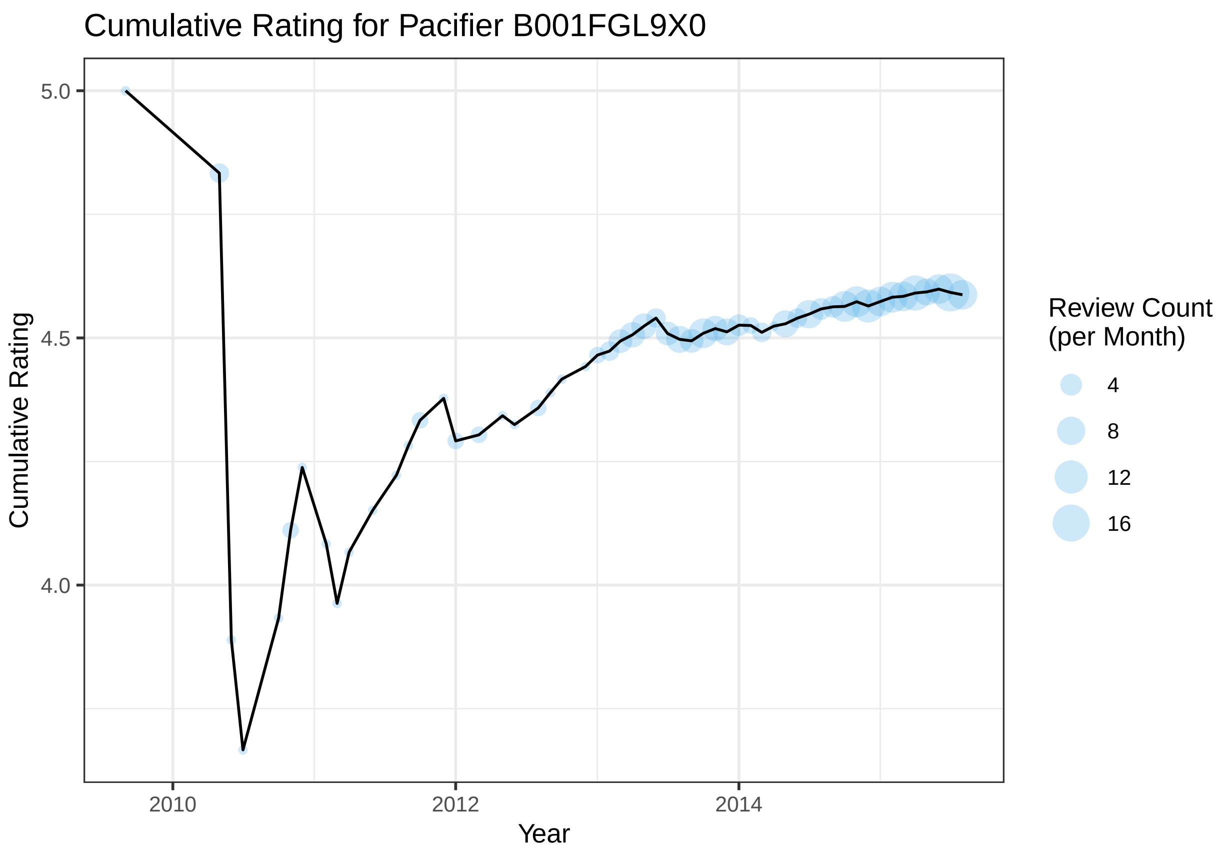


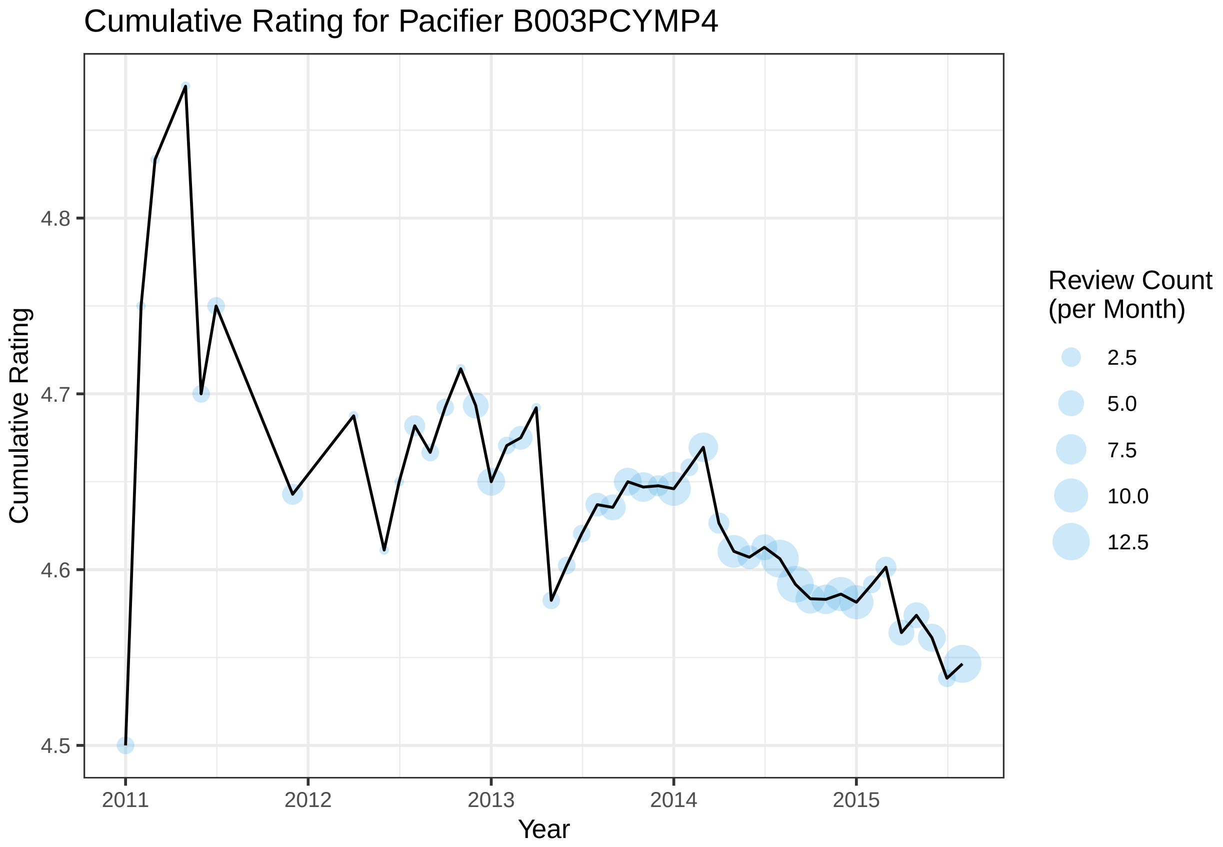
```
cum_rtg_plots(pop_paci,"#56B4E9",5,"Pacifier")
```


Cumulative Rating for Pacifier B003CK3LDI
Cumulative Rating
5.00
4.75
4.50
4.25
4.00
2011
2012
2013
2014
2015
Year
Review Count
(per Month)
5
10
15
20
25


Cumulative Rating for Pacifier B0028IDXDS
Cumulative Rating
5.0
4.8
4.6
4.4
4.2
2010
2012
2014
Year
Review Count
(per Month)
5
10
15


Cumulative Rating for Pacifier B0045I6IA4
Cumulative Rating
5.0
4.9
4.8
4.7
2012
2013
2014
2015
Year
Review Count
(per Month)
5
10
15
20


Cumulative Rating for Pacifier B001FGL9X0
Cumulative Rating
5.0
4.5
4.0
2010
2012
2014
Year
Review Count
(per Month)
4
8
12
16



6) Track number of reviews for given period after each review

Can use this method to compare dates

```
somedays <- dryer$review_date[c(1,20,30,50)]
somedayslim <- dryer$review_date[c(1,20,30,50)] + 7 # can use addition op + < to compare dates
somedays
```

```
## [1] "2015-08-31" "2015-08-29" "2015-08-28" "2015-08-26"
```

```
somedayslim
```

```
## [1] "2015-09-07" "2015-09-05" "2015-09-04" "2015-09-02"
```

```
somedays < somedayslim[3] & somedays > somedays[3] # checks if dates within 7 days of day 3
```

```
## [1]  TRUE  TRUE FALSE FALSE
```

6a) Look at review count for period after each star rating (boxplot)

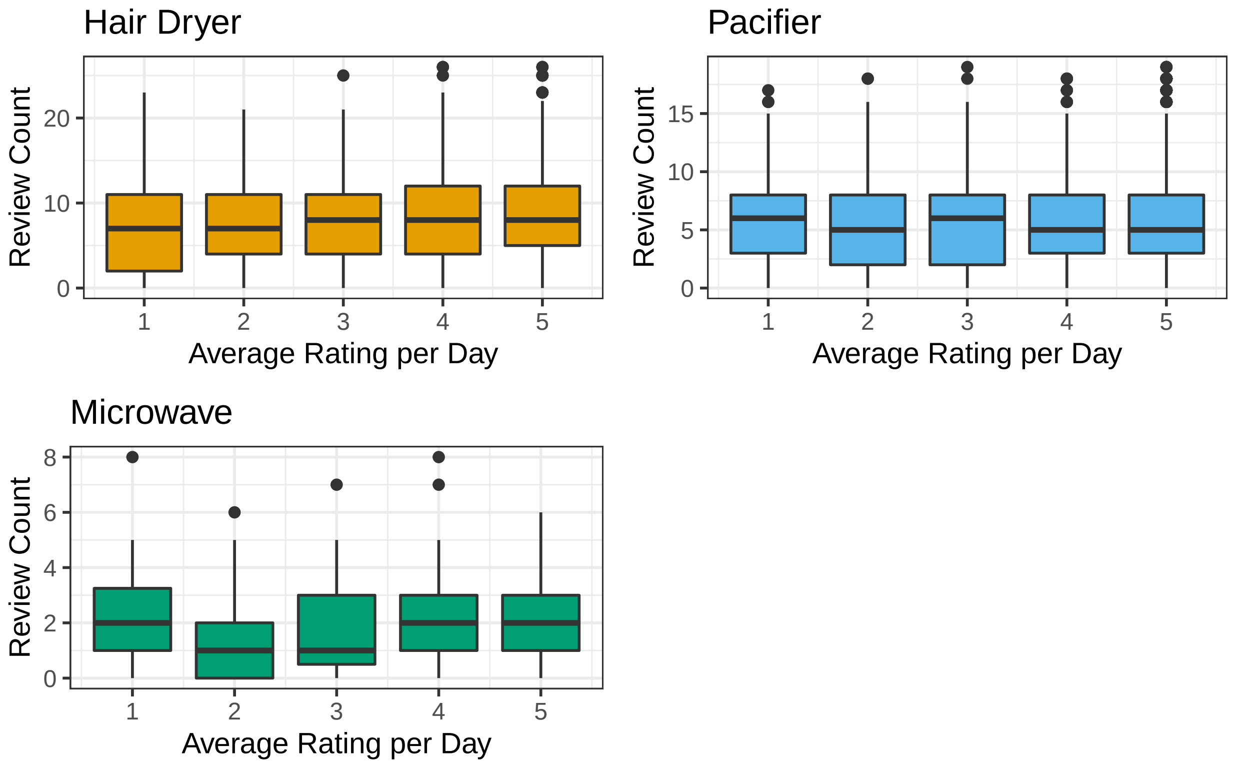
```
# will count the number of reviews after a review for a given period in days (default 3), delayed by th
reviews_after <- function(df,period=3,shipping=4){
  days <- df$review_date
  days_lower <- days + shipping
  days_upper <- days_lower + period
  reviews_after <- numeric(0)
  for(i in 1:length(days)){
    # checks how many reviews followed each review for [period] span after [shipping delay] period
    reviews_after <- append(reviews_after, sum(days < days_upper[i] & days > days_lower[i]))
  }
  df$reviews_after <- reviews_after
  df
}
```

```
dryer <- reviews_after(pop_dryer)
paci <- reviews_after(pop_paci)
micro <- reviews_after(pop_micro)
```

```
dryer_after <- dryer %>% group_by(review_date,product_id) %>% summarise(avg_rating = mean(star_rating),
                                                              avg_rating_rd = round(mean(star_rating)),
                                                              reviews_after = mean(reviews_after))
paci_after <- paci %>% group_by(review_date,product_id) %>% summarise(avg_rating = mean(star_rating),
                                                            avg_rating_rd = round(mean(star_rating)),
                                                            reviews_after = mean(reviews_after))
micro_after <- micro %>% group_by(review_date,product_id) %>% summarise(avg_rating = mean(star_rating),
                                                              avg_rating_rd = round(mean(star_rating)),
                                                              reviews_after = mean(reviews_after))
```

```
rr1 <- ggplot(dryer_after,aes(x=avg_rating_rd,y=reviews_after,group=avg_rating_rd)) + geom_boxplot(fill=
  theme_bw() + theme(legend.position = "none") +
  xlab("Average Rating per Day") + ylab("Review Count") +
  labs(title = "Hair Dryer")
rr2 <- ggplot(paci_after,aes(x=avg_rating_rd,y=reviews_after,group=avg_rating_rd)) + geom_boxplot(fill=
  theme_bw() + theme(legend.position = "none") +
  xlab("Average Rating per Day") + ylab("Review Count") +
  labs(title = "Pacifier")
rr3 <- ggplot(micro_after,aes(x=avg_rating_rd,y=reviews_after,group=avg_rating_rd)) + geom_boxplot(fill=
  theme_bw() + theme(legend.position = "none") +
  xlab("Average Rating per Day") + ylab("Review Count") +
  labs(title = "Microwave")


grid.arrange(rr1,rr2,rr3,ncol=2,top = "Review Count For 3 Day Period After (4 Days) Shipping for Each S
             bottom = "(Note: Each data point is a unique product on a unique day)")
```

Review Count For 3 Day Period After (4 Days) Shipping for Each Star Rating



(Note: Each data point is a unique product on a unique day)

```
dryer_after$avg_rating_rd <- as.factor(dryer_after$avg_rating_rd)
paci_after$avg_rating_rd <- as.factor(paci_after$avg_rating_rd)
micro_after$avg_rating_rd <- as.factor(micro_after$avg_rating_rd)
# dryer_after2
da_mod <- lm(reviews_after ~ avg_rating_rd,data = dryer_after)
pa_mod <- lm(reviews_after ~ avg_rating_rd,data = paci_after)
ma_mod <- lm(reviews_after ~ avg_rating_rd,data = micro_after)
cat("\nHair Dryer\n")
```

```
##
## Hair Dryer
```

```
anova(da_mod)
```

```
## Analysis of Variance Table
##
## Response: reviews_after
##                 Df Sum Sq Mean Sq F value    Pr(>F)
## avg_rating_rd    4    517 129.210  4.8629 0.0006515 ***
## Residuals     4551 120923  26.571
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```

```
cat("\nPacifier\n")
```

```
##
## Pacifier
```

```
anova(pa_mod)
```

```
## Analysis of Variance Table
##
## Response: reviews_after
##                  Df Sum Sq Mean Sq F value Pr(>F)
## avg_rating_rd    4     22  5.4135  0.3224 0.8631
## Residuals     2904  48761 16.7909
```

```
cat("\nMicrowave\n")
```

```
##
## Microwave
```

```
anova(ma_mod)
```

```
## Analysis of Variance Table
##
## Response: reviews_after
##                 Df Sum Sq Mean Sq F value Pr(>F)
## avg_rating_rd   4   6.16  1.5410  0.5218 0.7198
## Residuals     204 602.45  2.9532
```

```
# will count the number of reviews after a review for a given period in days (default 3)
reviews_after <- function(df,days_after=3){
  days <- df$review_date
  days_upper <- df$review_date + days_after
  reviews_after <- numeric(0)
  for(i in 1:length(days)){
    # checks how many reviews followed each review for [day] span
    reviews_after <- append(reviews_after, sum(days < days_upper[i] & days > days[i]))
  }
  df$reviews_after_3 <- reviews_after
  df
}
```

6c) look at avg star rating of reviews after each star rating (does bad rating indicate more bad ratings to come?) boxplot?

```
# will look at avg rating of next [next_r] reviews, default 3
ratings_after <- function(df,next_r=3){

  df <- df %>% arrange(review_date) # sort so earliest date first
  avg_rating_after <- numeric(0)

  for(i in 1:nrow(df)){
    # find average rating for next [next_r] reviews
    if(i + next_r <= nrow(df)){
      rating_after <- mean(df$star_rating[(i+1):(i+next_r)])
    }
    else{
      rating_after <- NaN
    }
    avg_rating_after <- append(avg_rating_after, rating_after)
  }
  df$avg_rating_after <- avg_rating_after
  df
}
```

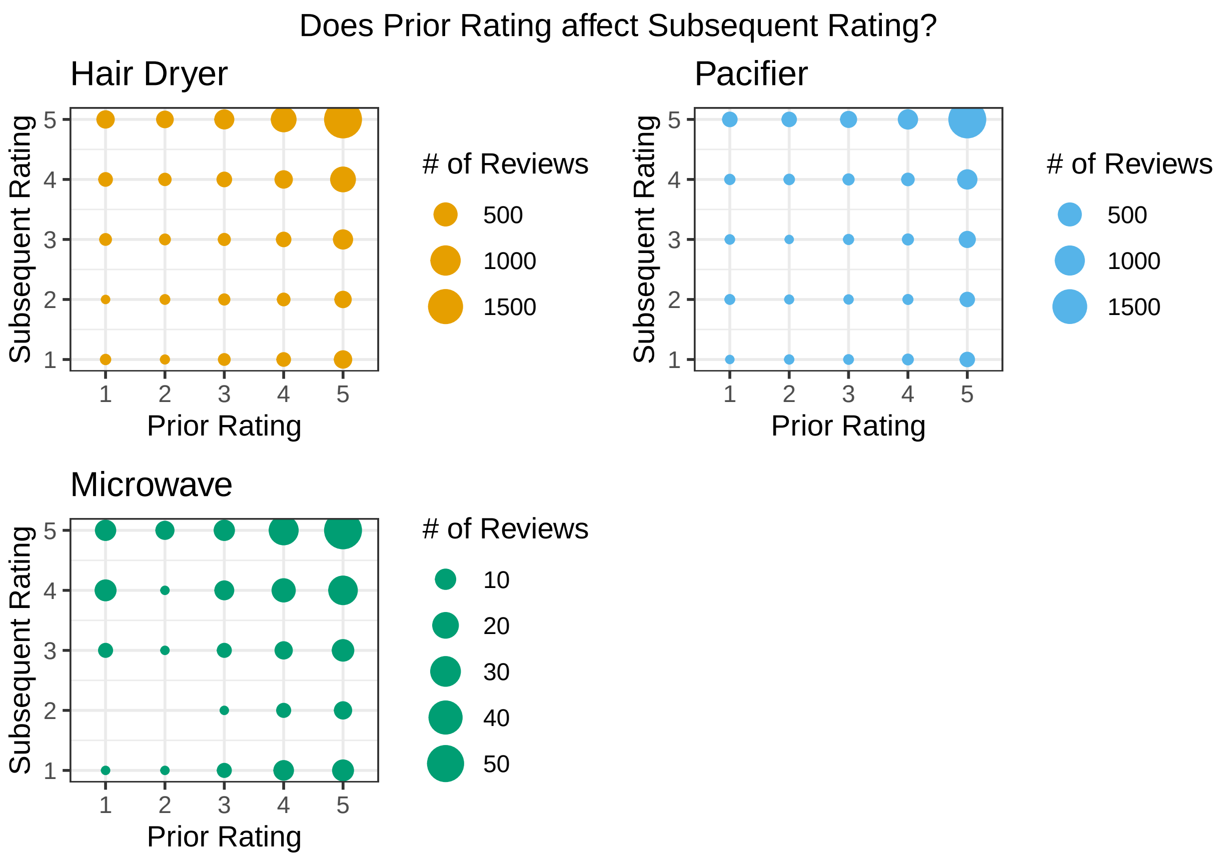
```
dryer_after3 <- ratings_after(pop_dryer,1)
paci_after3 <- ratings_after(pop_paci,1)
micro_after3 <- ratings_after(pop_micro,1)
dryer_after3
```

```
## # A tibble: 5,205 x 26
## # Groups:   product_id [24]
##    marketplace customer_id review_id product_id product_parent product_title
##    <chr>             <dbl> <chr>     <chr>               <dbl> <chr>
##  1 us             43740490 R2XM83JY~ B00005OOMZ      694290590 conair corp ~
##  2 us             37733836 R3GO6L5P~ B00005OOMZ      694290590 conair corp ~
##  3 us             50473837 R12APPEF~ B00005OOMZ      694290590 conair corp ~
##  4 us             51785663 R3NZ6I1E~ B0002G214U      685652978 conair soft ~
##  5 us             38641465 R37KYGDK~ B0009XH6TG       47684938 andis 1875-w~
##  6 us             52683833 RU2ONJLY~ B0009XH6TG       47684938 andis 1875-w~
##  7 us             51675091 RXK8FS5P~ B0008ENT8I      868768702 proversa jwm~
##  8 us             44499311 R3HEA8B7~ B0008ENT8I      868768702 proversa jwm~
##  9 us             35647799 R17YDVO1~ B00005OOMZ      694290590 conair corp ~
## 10 us             50896876 R2TVLB5Z~ B0009XH6TG       47684938 andis 1875-w~
## # ... with 5,195 more rows, and 20 more variables: product_category <chr>,
## #   star_rating <dbl>, helpful_votes <dbl>, total_votes <dbl>, vine <chr>,
## #   verified_purchase <chr>, review_headline <chr>, review_body <chr>,
## #   review_date <date>, type <chr>, helpful_ratio <dbl>, helpful <lgl>,
## #   has_votes <lgl>, impact_pos <dbl>, impact_neg <dbl>, cum_star_rating <dbl>,
## #   review_month <date>, impact_star <chr>, review_count <int>,
## #   avg_rating_after <dbl>
```

5s lead to more 5s

```
raa1 <- ggplot(dryer_after3,aes(x=factor(star_rating),y=avg_rating_after)) + geom_count(color=cbp[2]) +
  labs(size="# of Reviews",x="Prior Rating",y="Subsequent Rating") + theme_bw()
raa2 <- ggplot(paci_after3,aes(x=factor(star_rating),y=avg_rating_after)) + geom_count(color=cbp[3]) + g
   labs(size="# of Reviews",x="Prior Rating",y="Subsequent Rating") + theme_bw()
raa3 <- ggplot(micro_after3,aes(x=factor(star_rating),y=avg_rating_after)) + geom_count(color=cbp[4]) +
   labs(size="# of Reviews",x="Prior Rating",y="Subsequent Rating") + theme_bw()

grid.arrange(raa1,raa2,raa3,ncol=2,top = "Does Prior Rating affect Subsequent Rating?")
```

```
## Warning: Removed 1 rows containing non-finite values (stat_sum).

## Warning: Removed 1 rows containing non-finite values (stat_sum).

## Warning: Removed 1 rows containing non-finite values (stat_sum).
```



[NOT USED]

```
# ggplot(dryer_after3,aes(x=vine,y=avg_rating_after)) + geom_count(color=cbp[2]) + ggtitle("Hair Dryer"
  # labs(size="# of Reviews",x="Rating",y="Subsequent Rating") + theme_bw()
```

[NOT USED]

```
reviews_after2 <- function(df,num,type){
  # glist <- vector(mode = "list", length = num)
  i <- 1
  for(product in unique(df$product_id)[1:num]){
    prod_df <- df[df$product_id == product,]
    g <- ggplot(prod_df,aes(x=impactfuls,y=reviews_after,fill=impactfuls)) + geom_boxplot() +
      scale_fill_manual(values=cbp[c(7,6)]) + theme_bw() + theme(legend.position = "none") +
      xlab("Helpful Review that Day") + ylab("Review Count") +
      labs(title = paste("Review Count With vs. Without Helpful Reviews for",type,prod_df$product_id[1]
                   caption = "(Note: review count is for 3 days following initial review)")
    i <- i + 1
    plot(g)
  }
  # nCol <- floor(sqrt(num))
  # do.call("grid.arrange", c(glist, ncol=nCol))
}
```

[NOT USED]

```r
ra1 <- ggplot(dryer_after,aes(x=impactfuls,y=reviews_after,fill=impactfuls)) + geom_boxplot() +
  scale_fill_manual(values=cbp[c(7,6)]) + theme_bw() + theme(legend.position = "none") +
  xlab("Helpful Review that Day") + ylab("Review Count") +
  labs(title = "Review Count Following Days With or Without Helpful Reviews (Hair Dryers)",
             caption = "(Note: review count is for 3 days following initial review)")
ra2 <- ggplot(paci_after,aes(x=impactfuls,y=reviews_after,fill=impactfuls)) + geom_boxplot() +
  scale_fill_manual(values=cbp[c(7,6)]) + theme_bw() + theme(legend.position = "none") +
  xlab("Helpful Review that Day") + ylab("Review Count") +
  labs(title = "Review Count Following Days With or Without Helpful Reviews (Pacifiers)",
             caption = "(Note: review count is for 3 days following initial review)")
ra3 <- ggplot(micro_after,aes(x=impactfuls,y=reviews_after,fill=impactfuls)) + geom_boxplot() +
  scale_fill_manual(values=cbp[c(7,6)]) + theme_bw() + theme(legend.position = "none") +
  xlab("Helpful Review that Day") + ylab("Review Count") +
  labs(title = "Review Count Following Days With or Without Helpful Reviews (Microwaves)",
             caption = "(Note: review count is for 3 days following initial review)")

# ra1; ra2; ra3
```