FinalProject-DataWrangling

```
library(tidyverse)
```

```
## -- Attaching packages ---------------------------------------------------------------------------
```

```
## v ggplot2 3.3.0     v purrr   0.3.3
## v tibble  3.0.0     v dplyr   0.8.5
## v tidyr   1.0.3     v stringr 1.4.0
## v readr   1.3.1     v forcats 0.5.0
```

```
## -- Conflicts ------------------------------------------------------------------------------------
## x dplyr::filter() masks stats::filter()
## x dplyr::lag()    masks stats::lag()
```

```
library(lubridate)
```

```
##
## Attaching package: 'lubridate'
```

```
## The following objects are masked from 'package:dplyr':
##
##     intersect, setdiff, union
```

```
## The following objects are masked from 'package:base':
##
##     date, intersect, setdiff, union
```

```
library(zoo)
```

```
##
## Attaching package: 'zoo'
```

```
## The following objects are masked from 'package:base':
##
##     as.Date, as.Date.numeric
```

```
library(gridExtra) # for plotting side by side
```

```
##
## Attaching package: 'gridExtra'
```

```
## The following object is masked from 'package:dplyr':
##
##     combine
```

```
library(scales) # for previewing colors
```

```
##
## Attaching package: 'scales'
```

```
## The following object is masked from 'package:purrr':
##
##     discard
```

```
## The following object is masked from 'package:readr':
##
##     col_factor
```

```
# paci <- read_tsv("pacifier.tsv")
# micro <- read_tsv("microwave.tsv")
# # head(micro)
# dryer <- read_tsv("hair_dryer.tsv")
# # head(dryer)
# glimpse(paci)
# glimpse(micro)
# glimpse(dryer)
```

```
dryer <- read_csv("hair_dryer.csv")
paci <- read_csv("pacifier.csv")
micro <- read_csv("microwave.csv")
```

1) Dates should be a datetime format for comparison

```
paci$review_date <- as.Date(paci$review_date, "%m/%d/%Y")
dryer$review_date <- as.Date(dryer$review_date, "%m/%d/%Y")
micro$review_date <- as.Date(micro$review_date, "%m/%d/%Y")
head(micro)
```

```
## # A tibble: 6 x 24
##   marketplace customer_id review_id product_id product_parent product_title
##   <chr>             <dbl> <chr>     <chr>               <dbl> <chr>
## 1 us             21879631 RY52KZAB~ B0052G14E8      423421857 danby 0.7 cu~
## 2 us             14964566 R3GCOEV4~ B0055UBB4O      423421857 danby 0.7 cu~
## 3 us             13230389 R1V2OPPN~ B0052G14E8      423421857 danby 0.7 cu~
## 4 us             43655888 R9QOQDTL~ B004ZU09QQ      423421857 danby 0.7 cu~
## 5 us               117794 R3DL7HYC~ B005GSZB7I      827502283 whirlpool st~
## 6 us             16018452 R3M88678~ B004ZU09QQ      423421857 danby 0.7 cu~
## # ... with 18 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>
```

2) Fix incosistent capitalization in pacifier and microwave data -> make uniform

```
dryer$marketplace <- tolower(dryer$marketplace)
paci$marketplace <- tolower(paci$marketplace)
paci$product_category <- tolower(paci$product_category)
micro$marketplace <- tolower(micro$marketplace)
micro$product_category <- tolower(micro$product_category)
```

do same for review headline and body

```
paci$review_headline <- tolower(paci$review_headline)
paci$review_body <- tolower(paci$review_body)
micro$review_headline <- tolower(micro$review_headline)
```

```
micro$review_body <- tolower(micro$review_body)
dryer$review_headline <- tolower(dryer$review_headline)
dryer$review_body <- tolower(dryer$review_body)
```

and vine + verified

```
paci$vine <- tolower(paci$vine)
paci$verified_purchase <- tolower(paci$verified_purchase)
micro$vine <- tolower(micro$vine)
micro$verified_purchase <- tolower(micro$verified_purchase)
dryer$vine <- tolower(dryer$vine)
dryer$verified_purchase <- tolower(dryer$verified_purchase)
```

3a) Create helpful votes/total votes ratio column (lots of 0s and 1s, meaning this isn't best metric)

```
helpful_ratio <- function(df){
  df$helpful_ratio <- df$helpful_votes / df$total_votes
  return(df)
}
```

```
dryer <- helpful_ratio(dryer)
paci <- helpful_ratio(paci)
micro <- helpful_ratio(micro)
```

3b) Create indicator for whether review has any votes

```
vote_indicator <- function(df){
  df$has_votes <- df$total_votes > 0
  return(df)
}
```

```
dryer <- vote_indicator(dryer)
paci <- vote_indicator(paci)
micro <- vote_indicator(micro)
```

3c) Create indicator for whether review was helpful or not

```
# no votes -> NA
# > 1 total votes and > 0.5 helful ratio -> TRUE
# else -> F
helpful_indicator <- function(df){
  dummy1 <- ifelse(df$total_votes > 1,T,NA) # reviews with 0 or 1 total votes are NA
  dummy2 <- df$helpful_ratio > 0.5 # stays NA if NA
  df$helpful <- dummy1 & dummy2
  return(df)
}
```

```
dryer <- helpful_indicator(dryer)
paci <- helpful_indicator(paci)
micro <- helpful_indicator(micro)
```

3d) Create an impact_pos and impact_neg column

```
# impact_pos indicates whehter a review is helpful and >= 4
is_impact_pos <- function(df){
  df$impact_pos <- ifelse(!(df$helpful %in% c(FALSE,NA)) & df$star_rating >= 4,1,0) # 1 = positive + im
  return(df)
}

# impact_neg indicates whehter a review is helpful and <= 3
is_impact_neg <- function(df){
  df$impact_neg<- ifelse(!(df$helpful %in% c(FALSE,NA)) & df$star_rating <= 3,1,0) # 1 = positive + imp
  return(df)
}
```

```
dryer <- is_impact_pos(dryer)
paci <- is_impact_pos(paci)
micro <- is_impact_pos(micro)
dryer <- is_impact_neg(dryer)
paci <- is_impact_neg(paci)
micro <- is_impact_neg(micro)
```

3e) Create impact star rating

```
impact_star <- function(df){
  df %>% mutate(impact_star = ifelse(helpful,paste("helpful",star_rating,"star"),NA))
}
```

```
dryer <- impact_star(dryer)
paci <- impact_star(paci)
micro <- impact_star(micro)
```

3f) Create a cumulative avg star rating

```
cum_rtg <- function(df){
  df %>% mutate(cum_star_rating = cummean(star_rating))
}
```

```
dryer <- cum_rtg(dryer)
paci <- cum_rtg(paci)
micro <- cum_rtg(micro)
```

3g) Create new review_month column

```
month_only <- function(df){
  df$review_month <- format(df$review_date, format="%Y-%m")
  df$review_month <- as.Date(as.yearmon(df$review_month, "%Y-%m"))
  df
}
```

```
dryer <- month_only(dryer)
paci <- month_only(paci)
micro <- month_only(micro)
```

4a) combine dfs into one for comparison purposes, adding type column

```
dryer$type <- "hair_dryer"
paci$type <- "pacifier"
micro$type <- "microwave"
all <- rbind(dryer,paci,micro)
```

```
write.csv(paci,"pacifier.csv", row.names = FALSE)
write.csv(micro,"microwave.csv", row.names = FALSE)
write.csv(dryer,"hair_dryer.csv", row.names = FALSE)
write.csv(all,"all_products.csv", row.names = FALSE)
```

4b) Filtered for verified users

```
paci_verified <- paci[paci$verified_purchase == "y",]
micro_verified <- micro[micro$verified_purchase == "y",]
dryer_verified <- dryer[dryer$verified_purchase == "y",]
all_verified <- all[all$verified_purchase == "y",]

write.csv(paci_verified,"pacifier_verif.csv", row.names = FALSE)
write.csv(micro_verified,"microwave_verif.csv", row.names = FALSE)
write.csv(dryer_verified,"hair_dryer_verif.csv", row.names = FALSE)
write.csv(all_verified,"all_verif.csv", row.names = FALSE)
```

4c) Filtered for vine reviews

```
paci_vine <- paci[paci$vine == "y",]
micro_vine <- micro[micro$vine == "y",]
dryer_vine <- dryer[dryer$vine == "y",]
all_vine <- all[all$vine == "y",]

write.csv(paci_vine,"pacifier_vine.csv", row.names = FALSE)
write.csv(micro_vine,"microwave_vine.csv", row.names = FALSE)
write.csv(dryer_vine,"hair_dryer_vine.csv", row.names = FALSE)
write.csv(all_vine,"all_vine.csv", row.names = FALSE)
```

5a) Group data by month and find monthly stats

Define custom color set

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

Find stats for each month

```
# combine titles and reviews for each month
monthly_stats <- function(df){
df_months <- df %>% group_by(review_month) %>% summarise(review_count = n(),
                                                         product_titles = paste0(product_title, collaps
                                                         review_headlines = paste0(review_headline, col
                                                         review_bodies = paste0(review_body, collapse =
                                                         star_ratings = paste0(star_rating, collapse =
                                                         avg_rating = mean(star_rating),
```

```
                                                    impact_pos = sum(impact_pos),
                                                    impact_neg = sum(impact_neg)
                                                    )
df_months <- df_months %>% mutate(impact_overall = impact_pos - impact_neg,cum_rating = cummean(avg_rati
df_months
}
```

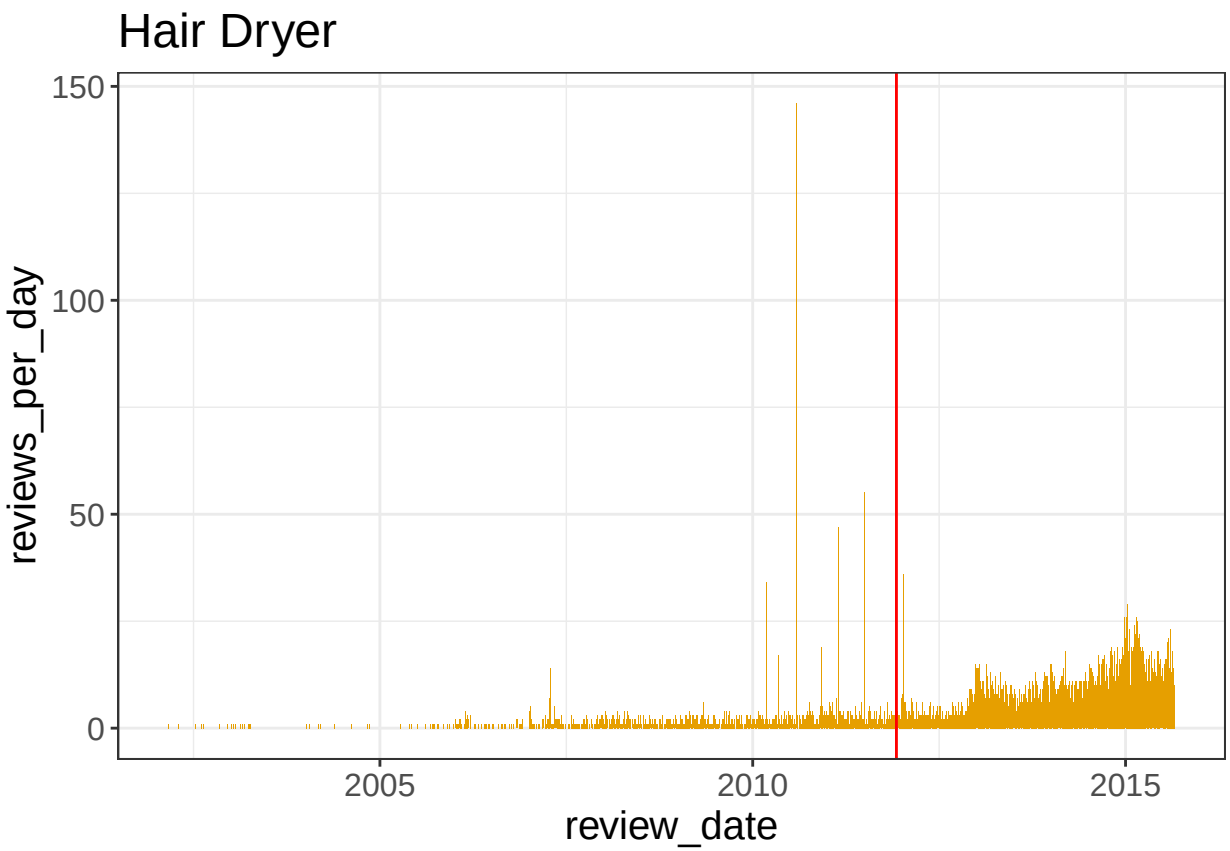
```
monthly_dryer <- monthly_stats(dryer)
monthly_paci <- monthly_stats(paci)
monthly_micro <- monthly_stats(micro)

# monthly_dryer
write.csv(monthly_paci,"monthly_paci.csv", row.names = FALSE)
write.csv(monthly_micro,"monthly_micro.csv", row.names = FALSE)
write.csv(monthly_dryer,"monthly_dryer.csv", row.names = FALSE)
```

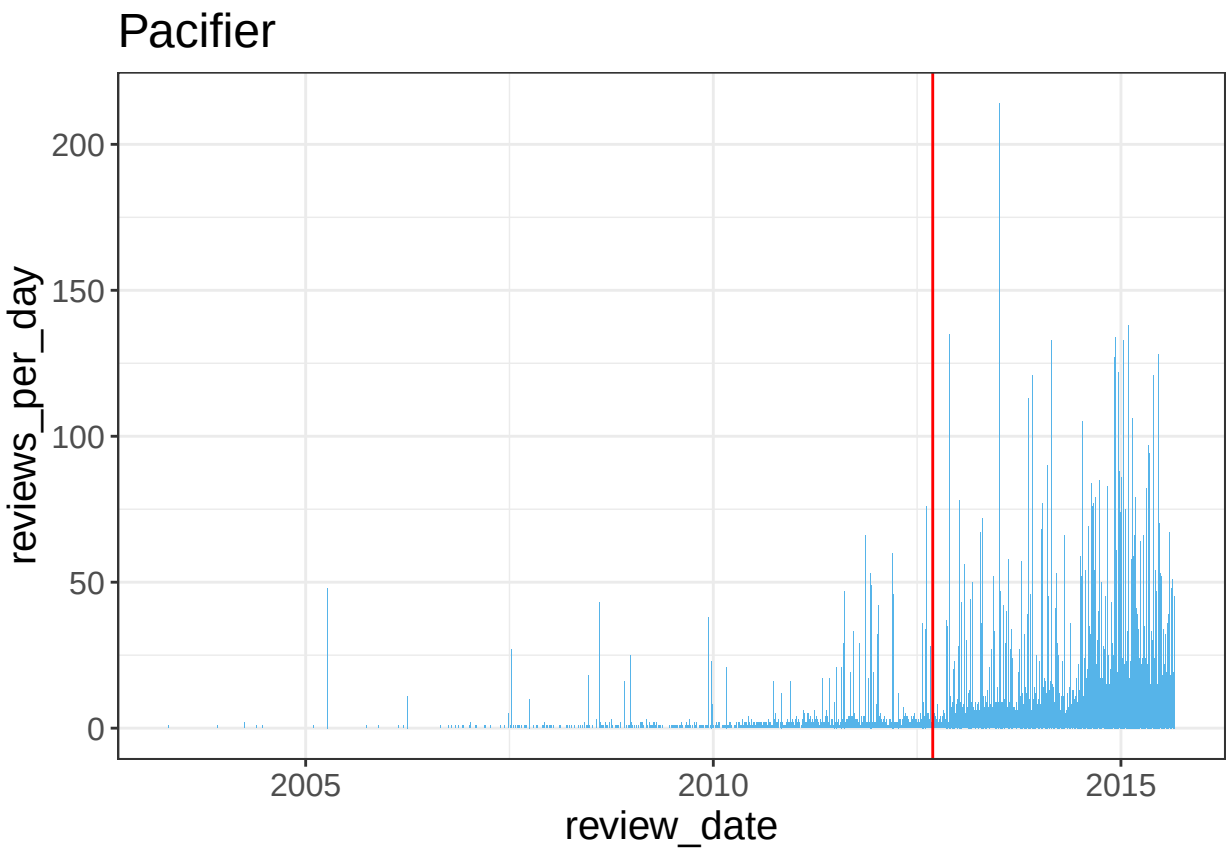
5b) Group data by day and find daily stats

First filter out years where not enough reviews

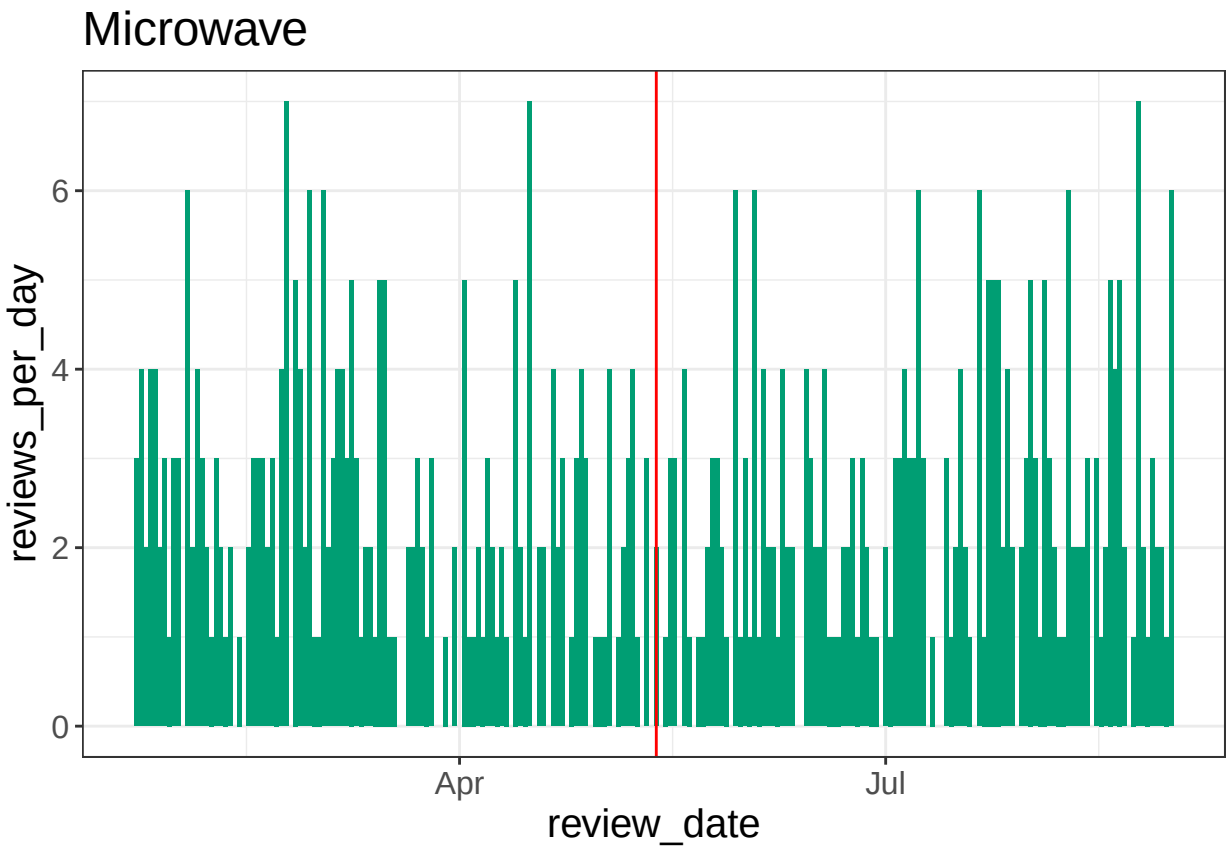
We can see that the bulk of the reviews for: Dryer are after 2010 -> filter fro days starting from 2010-03-11 (day with 34 reviews) Pacifier are after 2012 -> filter for days starting from 2011-12-04 (first day with > 50 reviews) Microwave are consistent at all times -> no filtering needed

```
dryer_daily <- dryer %>% group_by(review_date) %>% summarise(reviews_per_day = n())
paci_daily <- paci %>% group_by(review_date) %>% summarise(reviews_per_day = n())
micro_daily <- micro %>% group_by(review_date) %>% summarise(reviews_per_day = n())

dy1 <- ggplot(dryer_daily,aes(x=review_date,y=reviews_per_day)) + geom_col(fill=cbp[2]) +
  geom_vline(xintercept = mean(dryer_daily$review_date, na.rm = T), color = "red") +
  ggtitle("Hair Dryer") + theme_bw() + theme(text = element_text(size=15))
dy2 <- ggplot(paci_daily,aes(x=review_date,y=reviews_per_day)) + geom_col(fill=cbp[3]) +
  geom_vline(xintercept = mean(paci_daily$review_date, na.rm = T), color = "red") +
  ggtitle("Pacifier") + theme_bw() + theme(text = element_text(size=15))
dy3 <- ggplot(micro_daily,aes(x=review_date,y=reviews_per_day)) + geom_col(fill=cbp[4]) +
  geom_vline(xintercept = mean(micro_daily$review_date, na.rm = T), color = "red") +
  ggtitle("Microwave") + theme_bw() + theme(text = element_text(size=15))

# summary(dryer_daily$review_date)
# summary(paci_daily$review_date)
dy1; dy2; dy3
```


Hair Dryer
150
100
50
0
reviews_per_day
2005
2010
2015
review_date

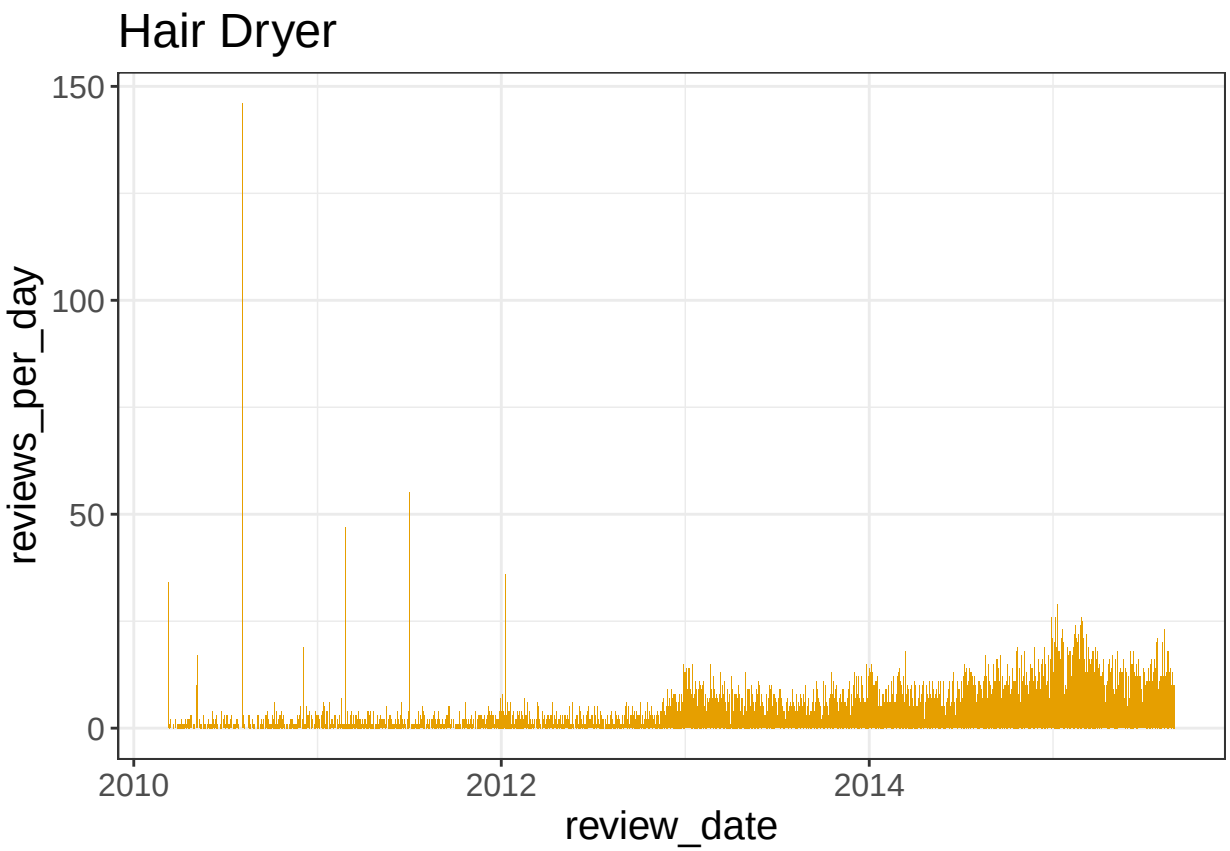


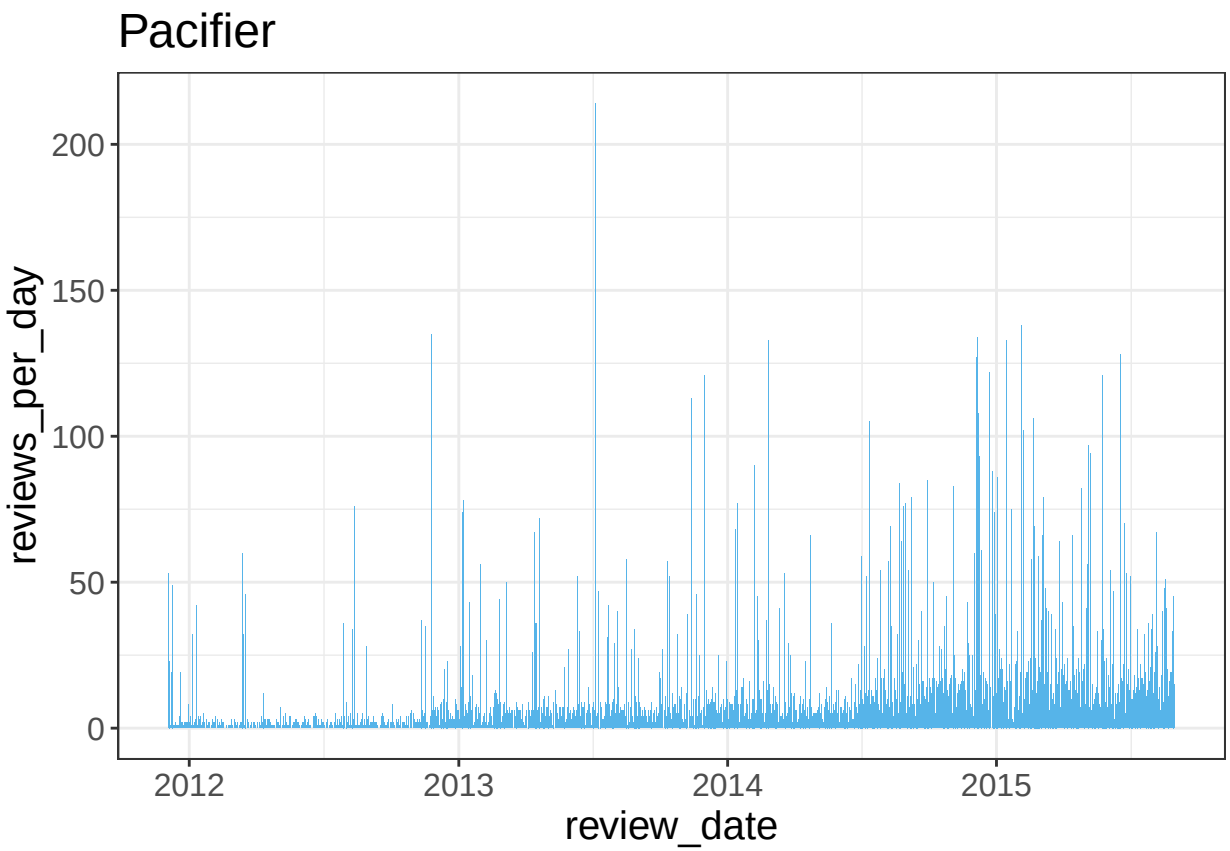

```
## Warning: Removed 1 rows containing missing values (position_stack).
```



Create new subsetted data -> reviews are much denser now

```
dryer2 <- dryer %>% filter(review_date >= "2010-03-11")
paci2 <- paci %>% filter(review_date >= "2011-12-04")
dryer_daily2 <- dryer2 %>% group_by(review_date) %>% summarise(reviews_per_day = n())
paci_daily2 <- paci2 %>% group_by(review_date) %>% summarise(reviews_per_day = n())

dy2a <- ggplot(dryer_daily2,aes(x=review_date,y=reviews_per_day)) + geom_col(fill=cbp[2]) +
  ggtitle("Hair Dryer") + theme_bw() + theme(text = element_text(size=15))
dy2b <- ggplot(paci_daily2,aes(x=review_date,y=reviews_per_day)) + geom_col(fill=cbp[3]) +
  ggtitle("Pacifier") + theme_bw() + theme(text = element_text(size=15))

dy2a; dy2b
```


Hair Dryer
150
100
50
0
reviews_per_day
2010
2012
2014
review_date



Then find stats for each day

```
# combine titles and reviews for each month
daily_stats <- function(df){
df_daily <- df %>% group_by(review_date) %>% summarise(review_count = n(),
                                                       product_titles = paste0(product_title, collapse =
                                                       review_headlines = paste0(review_headline, collap
                                                       review_bodies = paste0(review_body, collapse = "
                                                       star_ratings = paste0(star_rating, collapse = "
                                                       avg_rating = mean(star_rating),
                                                       impact_pos = sum(impact_pos),
                                                       impact_neg = sum(impact_neg)
                                                       )
df_daily <- df_daily %>% mutate(impact_overall = impact_pos - impact_neg, cum_rating = cummean(avg_ratin
df_daily
}
```

```
daily_dryer <- daily_stats(dryer2)
daily_paci <- daily_stats(paci2)
daily_micro <- daily_stats(micro)

write.csv(daily_paci,"daily_paci.csv", row.names = FALSE)
write.csv(daily_micro,"daily_micro.csv", row.names = FALSE)
write.csv(daily_dryer,"daily_dryer.csv", row.names = FALSE)
```