Tidy Tuesday - Week 10 (2021)

Julia Tache

3/05/2021

Superbowl Commercials

There are three things I look forward to when it comes to the Superbowl: finger foods, halftime show hot takes, and, of course, the commercials. The commercials arena is just as high stakes as the game itself: every year, companies pay literally millions of dollars to have their ads featured in the prime-time spot light. Some companies have become especially notorious for their Superbowl ads, including Doritos, who hosts a yearly contest encouraging creators to submit their original work.

Using data from superbowl-ads.com compiled by FiveThirtyEight, I explored some ad trends I found interesting. All of these data are from YouTube and include binary values on whether or not ads fit various themes, such as being "funny" or "patriotic." The fact that the data come from YouTube also helps measure how memorable these ads are- people usually choose to seek out videos on YouTube (or videos are recommended by the algorithm), so we can assume commercials that have higher view counts have a level of popularity that transcended their spot on TV. Longevity and impact are definitely useful when it comes to successful ad campaigns, and, hopefully, both of those factors translate to higher product sales.

```r
library(tidyverse)
library(stargazer)
tuesdata <- tidytuesdayR::tt_load('2021-03-02')
```

```
##
##  Downloading file 1 of 1: `youtube.csv`
```

```r
youtube <- tuesdata$youtube
head(youtube)
```

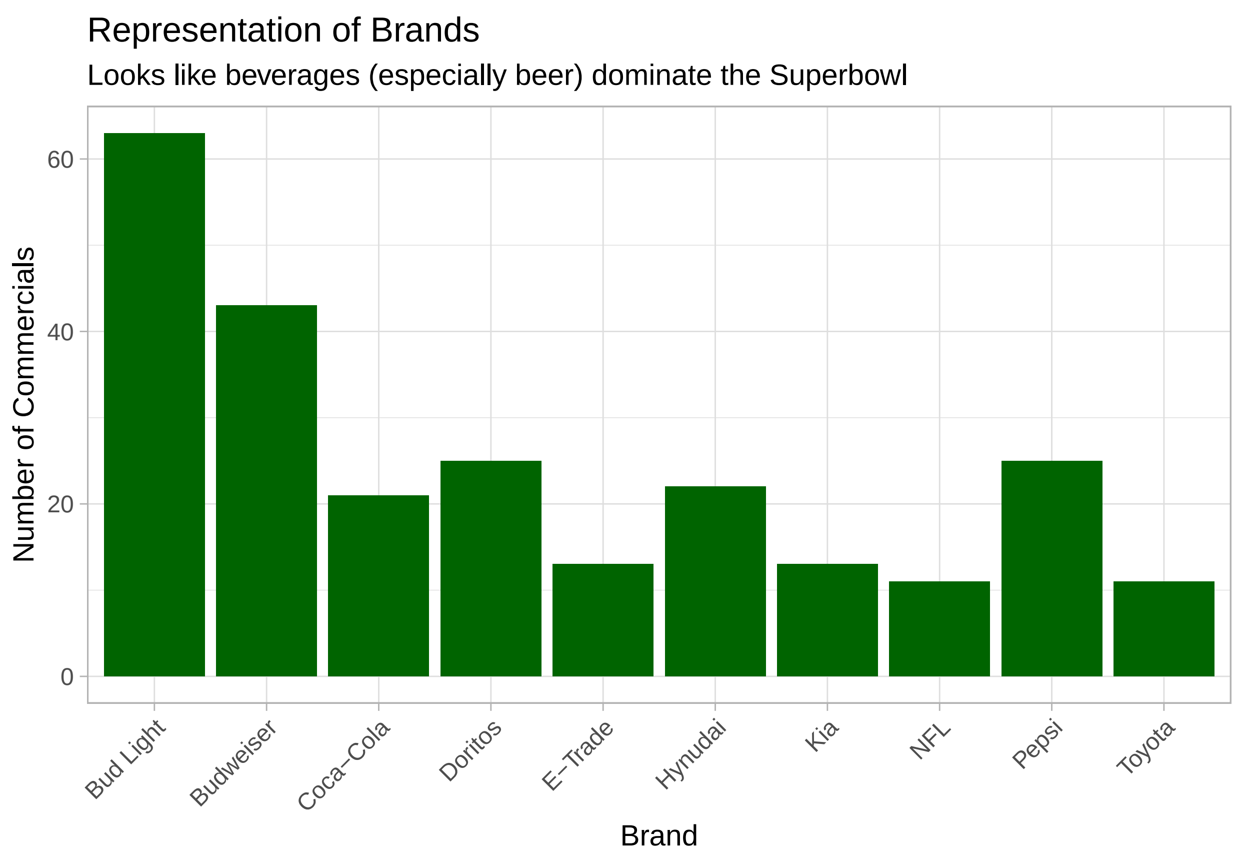
```
## # A tibble: 6 x 25
##    year brand superbowl_ads_d~ youtube_url funny show_product_qu~ patriotic
##   <dbl> <chr> <chr>            <chr>       <lgl> <lgl>            <lgl>
## 1  2018 Toyo~ https://superbo~ https://ww~ FALSE FALSE            FALSE
## 2  2020 Bud ~ https://superbo~ https://ww~ TRUE  TRUE             FALSE
## 3  2006 Bud ~ https://superbo~ https://ww~ TRUE  FALSE            FALSE
## 4  2018 Hynu~ https://superbo~ https://ww~ FALSE TRUE             FALSE
## 5  2003 Bud ~ https://superbo~ https://ww~ TRUE  TRUE             FALSE
## 6  2020 Toyo~ https://superbo~ https://ww~ TRUE  TRUE             FALSE
## # ... with 18 more variables: celebrity <lgl>, danger <lgl>,
## #   animals <lgl>, use_sex <lgl>, id <chr>, kind <chr>, etag <chr>,
## #   view_count <dbl>, like_count <dbl>, dislike_count <dbl>,
## #   favorite_count <dbl>, comment_count <dbl>, published_at <dttm>,
## #   title <chr>, description <chr>, thumbnail <chr>, channel_title <chr>,
## #   category_id <dbl>
```

Let's look at which brand dominates the Superbowl Commercial game.

```r
table(youtube$brand)
```

```
##
## Bud Light Budweiser Coca-Cola   Doritos   E-Trade   Hynudai       Kia
##        63        43        21        25        13        22        13
##       NFL     Pepsi    Toyota
##        11        25        11
```

```r
ggplot(youtube, aes(x = brand)) +
  geom_bar(fill = "dark green") +
  xlab("Brand") +
  ylab("Number of Commercials") +
  labs(title = "Representation of Brands",
       subtitle = "Looks like beverages (especially beer) dominate the Superbowl") +
  theme_light() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Budweiser's in the endzone, but what makes their ads truly touch-down worthy? Does their prevelance translate into views?

(also, excuse my terrible football puns, I know very little about the sport)

Next, I took a look at the most common words found in video descriptions:

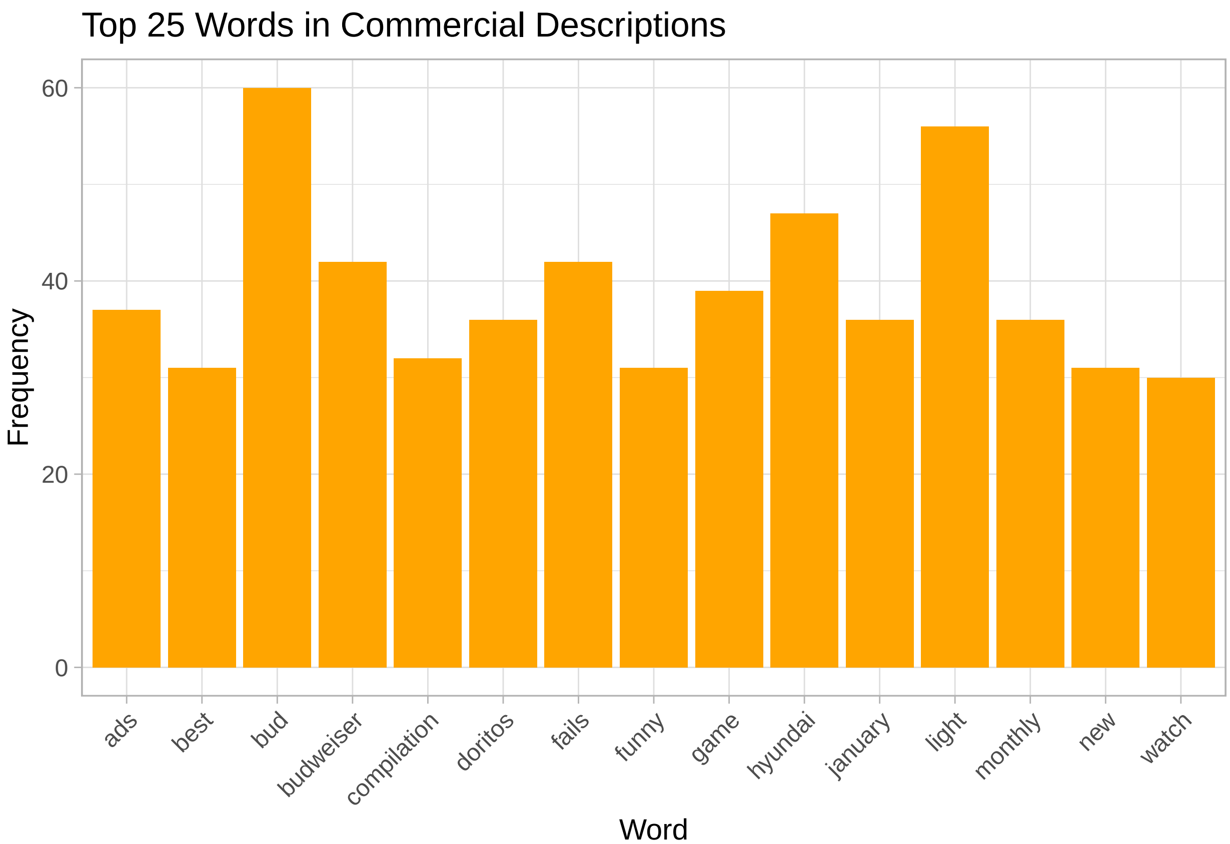
```r
library(tidytext)
library(tm)
library(SnowballC)
```

```r
library(RColorBrewer)
library(syuzhet)

TextDoc <- Corpus(VectorSource(youtube$description))

toSpace <- content_transformer(function (x , pattern ) gsub(pattern, " ", x))
TextDoc <- tm_map(TextDoc, toSpace, "/")
TextDoc <- tm_map(TextDoc, toSpace, "@")
TextDoc <- tm_map(TextDoc, toSpace, "\\|")
TextDoc <- tm_map(TextDoc, toSpace, "www")
TextDoc <- tm_map(TextDoc, toSpace, "http|https")
TextDoc <- tm_map(TextDoc, toSpace, ".com$")
TextDoc <- tm_map(TextDoc, toSpace, "youtube")
TextDoc <- tm_map(TextDoc, toSpace, "nfl")
TextDoc <- tm_map(TextDoc, toSpace, "commercial(s)")
TextDoc <- tm_map(TextDoc, toSpace, "bowl")
TextDoc <- tm_map(TextDoc, toSpace, "super")

# Convert the text to lower case
TextDoc <- tm_map(TextDoc, content_transformer(tolower))

# Remove numbers
TextDoc <- tm_map(TextDoc, removeNumbers)

# Remove english common stopwords
TextDoc <- tm_map(TextDoc, removeWords, stopwords("english"))

# Remove punctuations
TextDoc <- tm_map(TextDoc, removePunctuation)

# Eliminate extra white spaces
TextDoc <- tm_map(TextDoc, stripWhitespace)

# Use wordstems
TextDoc <- tm_map(TextDoc, wordStem)
```

```r
# Build a term-document matrix
TextDoc_dtm <- TermDocumentMatrix(TextDoc)
dtm_m <- as.matrix(TextDoc_dtm)

# Sort by descearing value of frequency
dtm_v <- sort(rowSums(dtm_m),decreasing=TRUE)
dtm_d <- data.frame(word = names(dtm_v),freq=dtm_v)
dtm_d <- dtm_d %>%
  rename("Word" = "word") %>%
  rename("Frequency" = "freq")

top25 <- head(dtm_d, 25)[6:20, ]
```

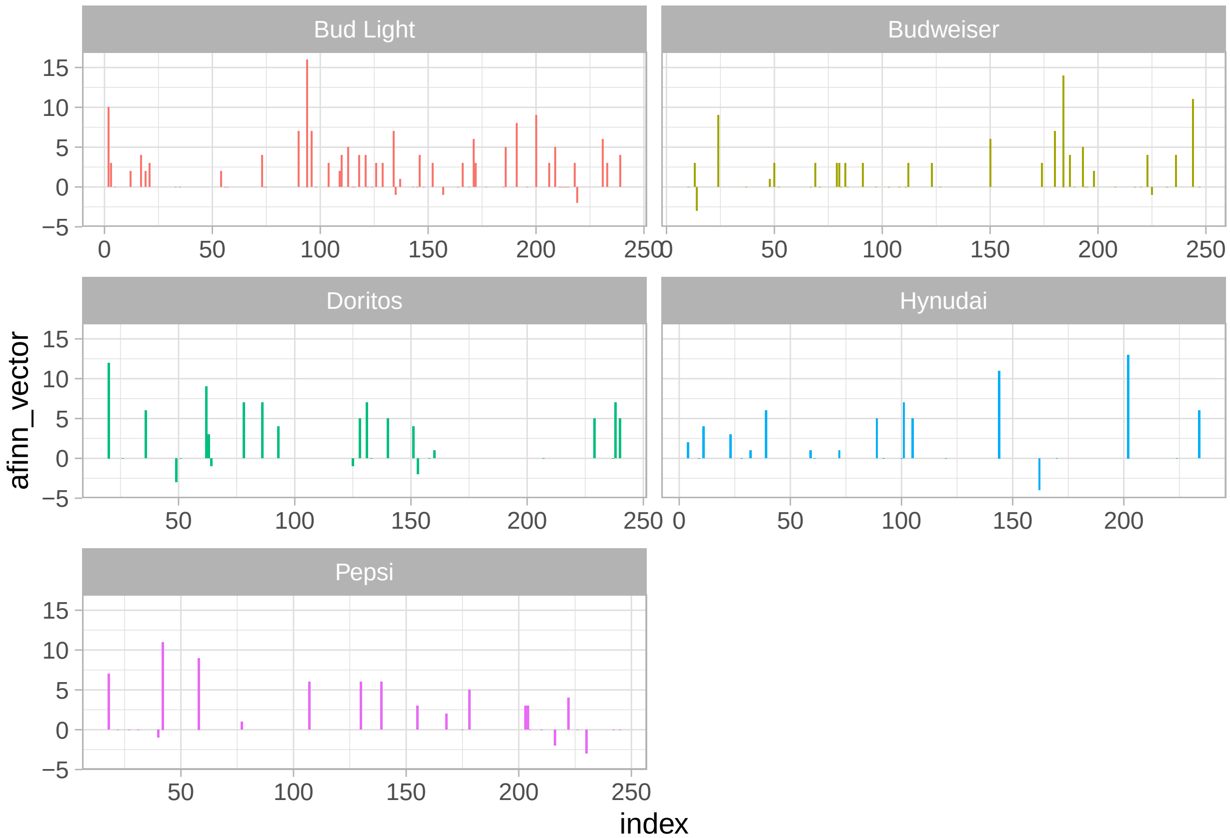
```r
top25 <- as.data.frame(top25)

ggplot(top25, aes(x = Word, y = Frequency)) +
  geom_bar(stat = "identity", fill = "orange") +
```

```r
  labs(title = "Top 25 Words in Commercial Descriptions") +
  theme_light() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Pretty much what one would expect for the MVPs: game, funny, doritos, etc. I then calculated the AFINN sentiment score for the description:

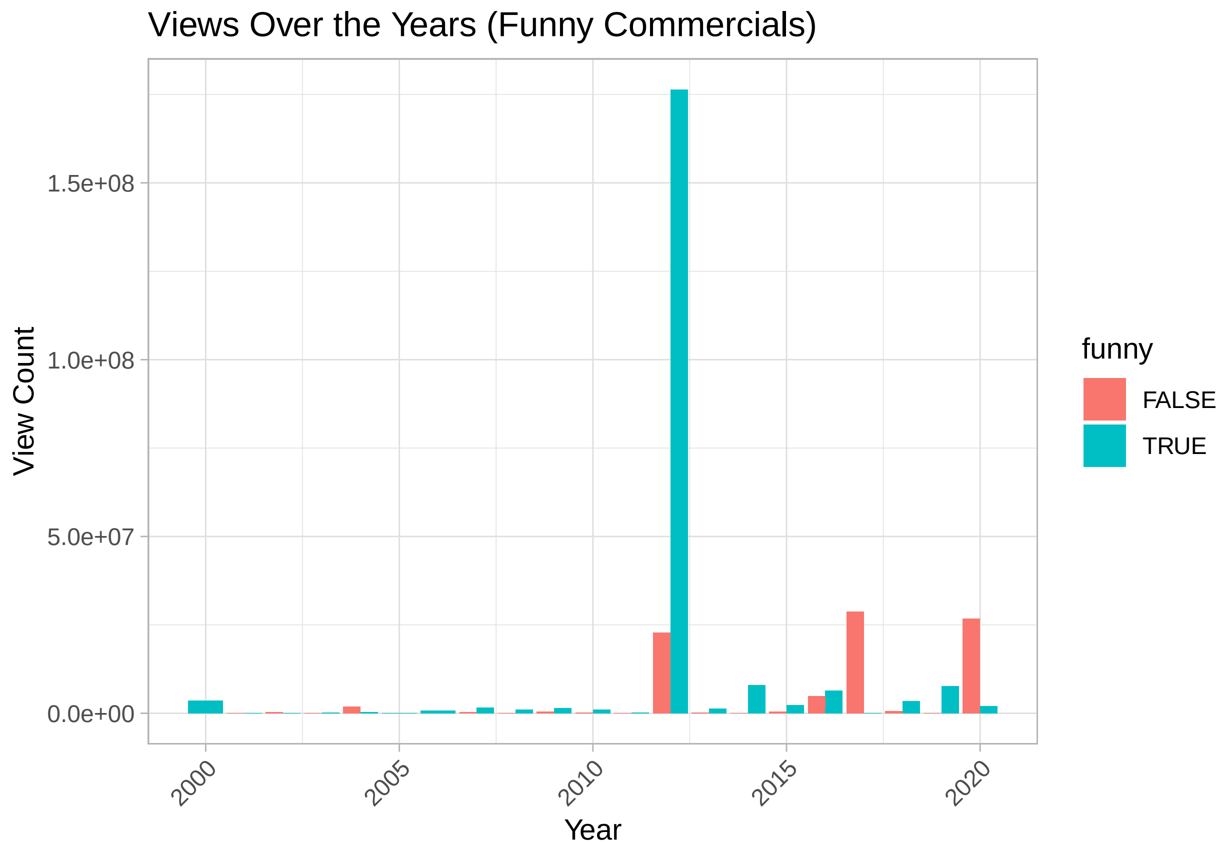
```r
youtube$clean_description <- str_remove(youtube$description, "[[:punct:]]")
youtube$clean_description <- str_replace(youtube$clean_description, "\n", " ")
youtube$clean_description <- str_remove(youtube$clean_description, stopwords("english"))
youtube$clean_description <- tolower(youtube$clean_description)
youtube$clean_description <- gsub(" ?(f|ht)tp(s?)://(.*)[.][a-z]+?(/[a-z]+/)?([alnum]+)", "", youtube$cl
youtube$clean_description <- str_remove(youtube$clean_description, "\\d{1,}")
youtube$clean_description <- str_split(youtube$clean_description, boundary("word")) # tokenize

youtube$clean_description <- as.character(youtube$clean_description)

youtube$afinn_vector <- get_sentiment(youtube$clean_description, method="afinn")

youtube_top5 <- youtube %>%
  mutate(index = row_number()) %>%
  filter(brand == "Bud Light" | brand == "Budweiser" | brand == "Pepsi" | brand == "Doritos" | brand ==

ggplot(youtube_top5, aes(index, afinn_vector, fill = brand)) +
  geom_col(show.legend = FALSE) +
  facet_wrap(~brand, ncol = 2, scales = "free_x") +
  theme_light()
```



Popular brands have overwhemingly positive descriptions, solidifying their place as GOATs.

So, what makes a commercial successful? When I think of Superbowl ads, my personal favorites are the funny ones.

```
ggplot(youtube, aes(x = year, y = view_count, fill = funny)) +
  geom_bar(stat = "identity", position = "dodge") +
  labs(title = "Views Over the Years (Funny Commercials)",
       x = "Year",
       y = "View Count") +
  theme_light() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



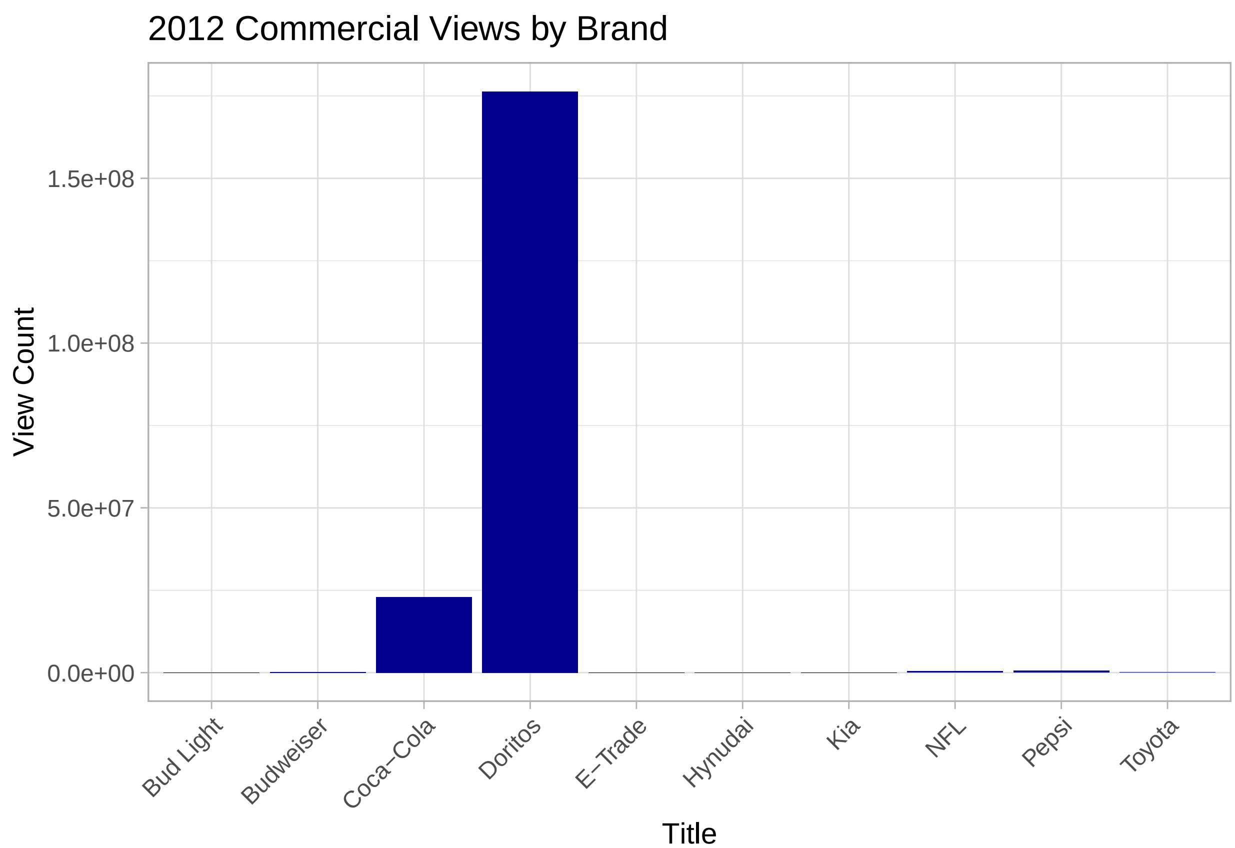
There's a huge spike in "funny" clips - and views overall - post 2010. After 2015, it seems like things got a bit more serious.

```
view_count_df <- youtube %>%
  group_by(year) %>%
  summarize(avg_view_count = mean(view_count))

view_count_df[order(-view_count_df$avg_view_count), ]
```

```
## # A tibble: 21 x 2
##     year avg_view_count
##    <dbl>          <dbl>
##  1  2012      13383538
##  2  2017       6009988
##  3  2020       3242159.
##  4  2007        305519.
##  5  2008        293292
##  6  2009        267306.
##  7  2004        258332
##  8  2006        180834.
##  9  2013        159624.
## 10  2011         62803.
## # ... with 11 more rows
```

Looks like there's a big spike in views in videos from 2012, so let's dig deeper.

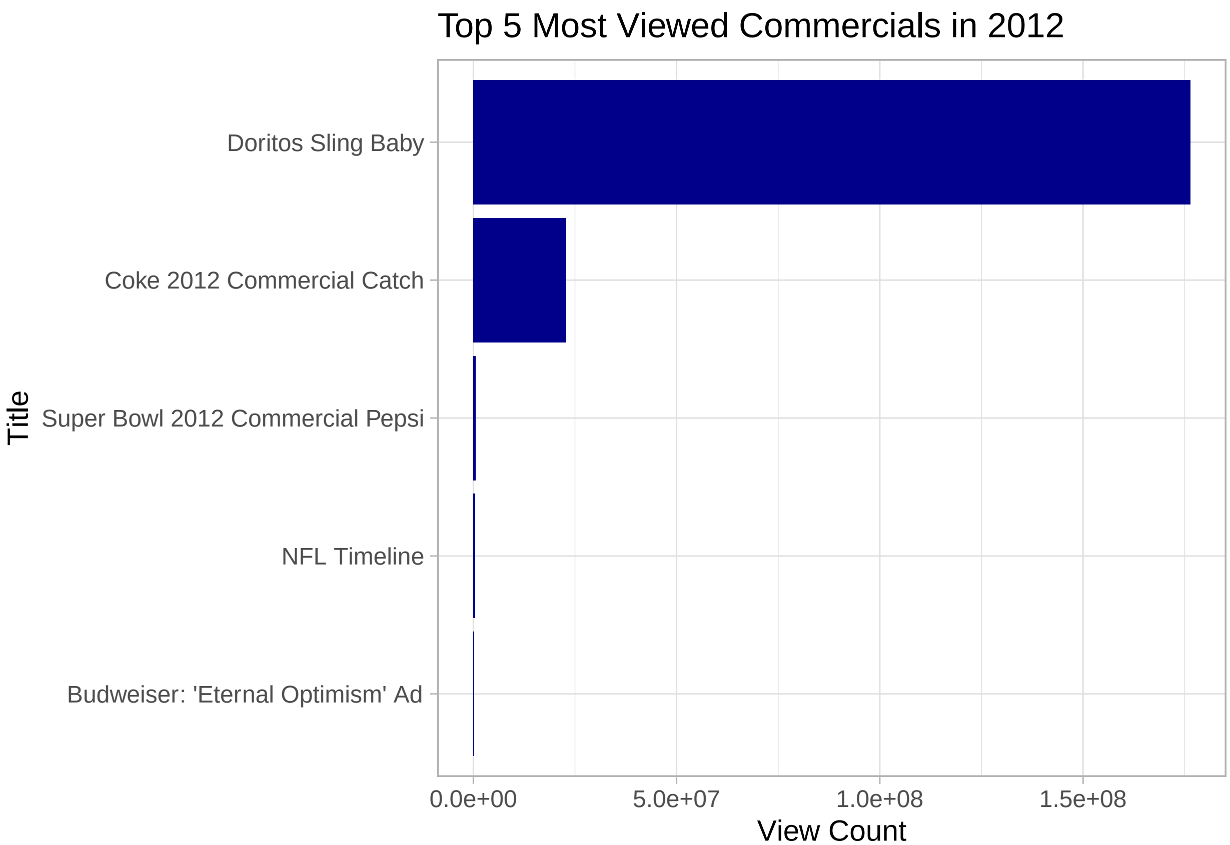
```r
yt_2012 <- youtube %>%
  filter(year == 2012)

ggplot(yt_2012, aes(x = brand, y = view_count)) +
  geom_bar(stat = "identity", position = "dodge", fill = "dark blue") +
  labs(title = "2012 Commercial Views by Brand",
       x = "Title",
       y = "View Count") +
  theme_light() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



Doritos was severely tackling the rest of the brands at that time.

```r
top_5_2012 <- yt_2012[order(-yt_2012$view_count), ][1:5, ]

top_5_2012$title <- str_remove(top_5_2012$title, "[:punct:]")
top_5_2012$title <- str_remove(top_5_2012$title, '"')
top_5_2012$title <- str_remove(top_5_2012$title, '"')
top_5_2012$title <- str_remove(top_5_2012$title, " starring NE_Bear")
top_5_2012$title <- str_remove(top_5_2012$title, " - King's Court")
top_5_2012$title <- str_remove(top_5_2012$title, "The Cult  She Sells Sanctuary - ")

ggplot(top_5_2012, aes(x = reorder(title, view_count), y = view_count)) +
  geom_bar(stat = "identity", position = "dodge", fill = "dark blue") +
  coord_flip() +
  labs(title = "Top 5 Most Viewed Commercials in 2012",
       x = "Title",
```

```
       y = "View Count") +
  theme_light() +
  theme(axis.text.y = element_text(hjust = 1))
```

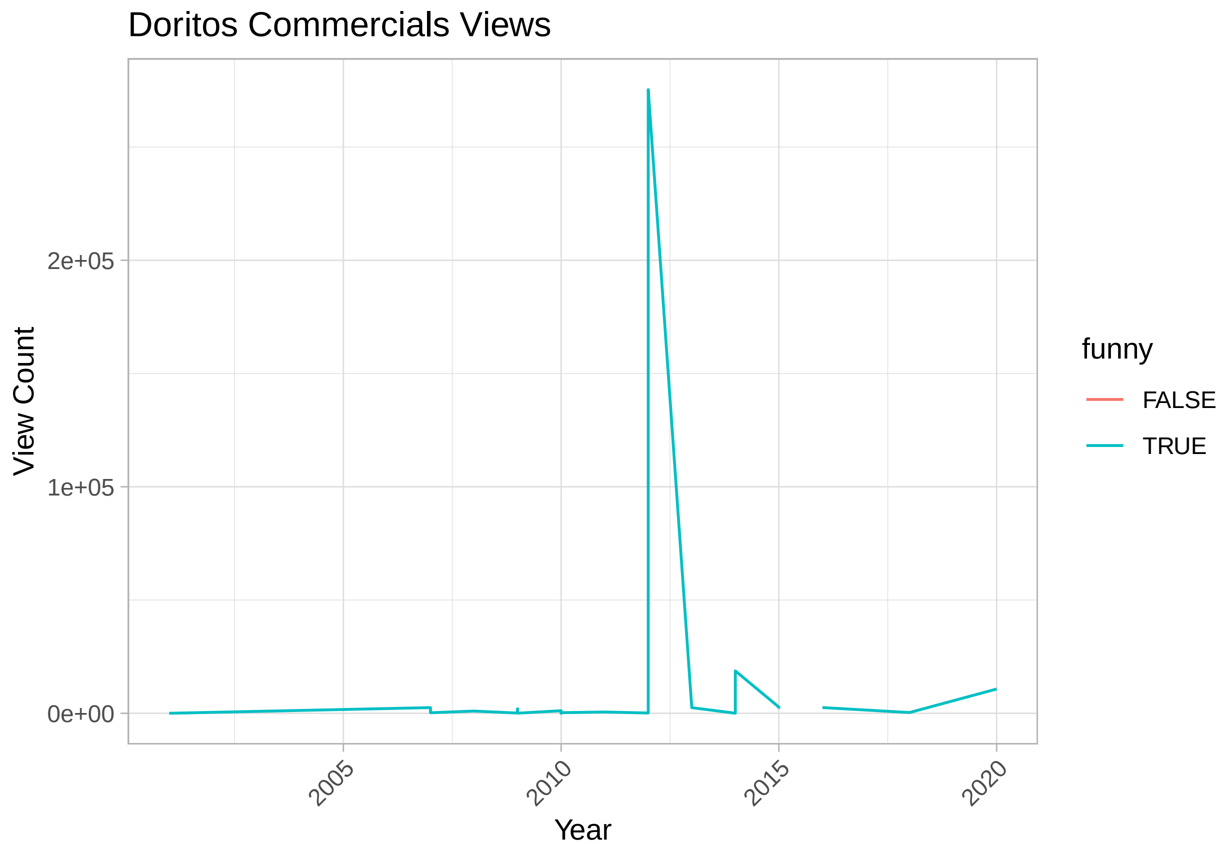


The most popular video is Sling Baby, which is quite funny: https://www.youtube.com/watch?v=6SWNLDdnz0A

(A personal note: I looked up the entity that produced this video, Madison McQueen, and found some…questionable political adverts made by the same creator. Proceed at your own risk.)

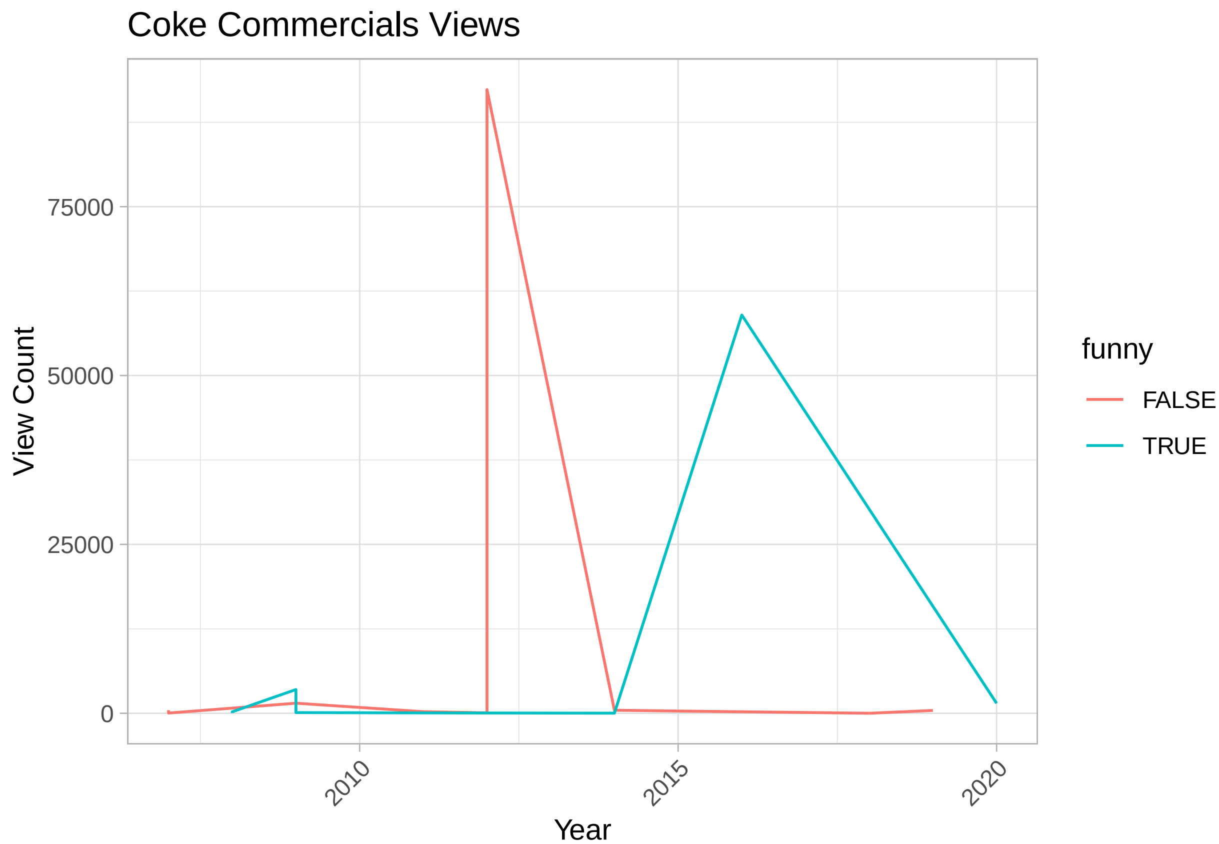
The graph below shows that Doritos main strategy over the years has been to be funny.

```
doritos <- youtube %>%
  filter(brand == "Doritos")

ggplot(doritos, aes(x = year, y = like_count, group = funny)) +
  geom_line(aes(color = funny)) +
  labs(title = "Doritos Commercials Views",
       x = "Year",
       y = "View Count") +
  theme_light() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



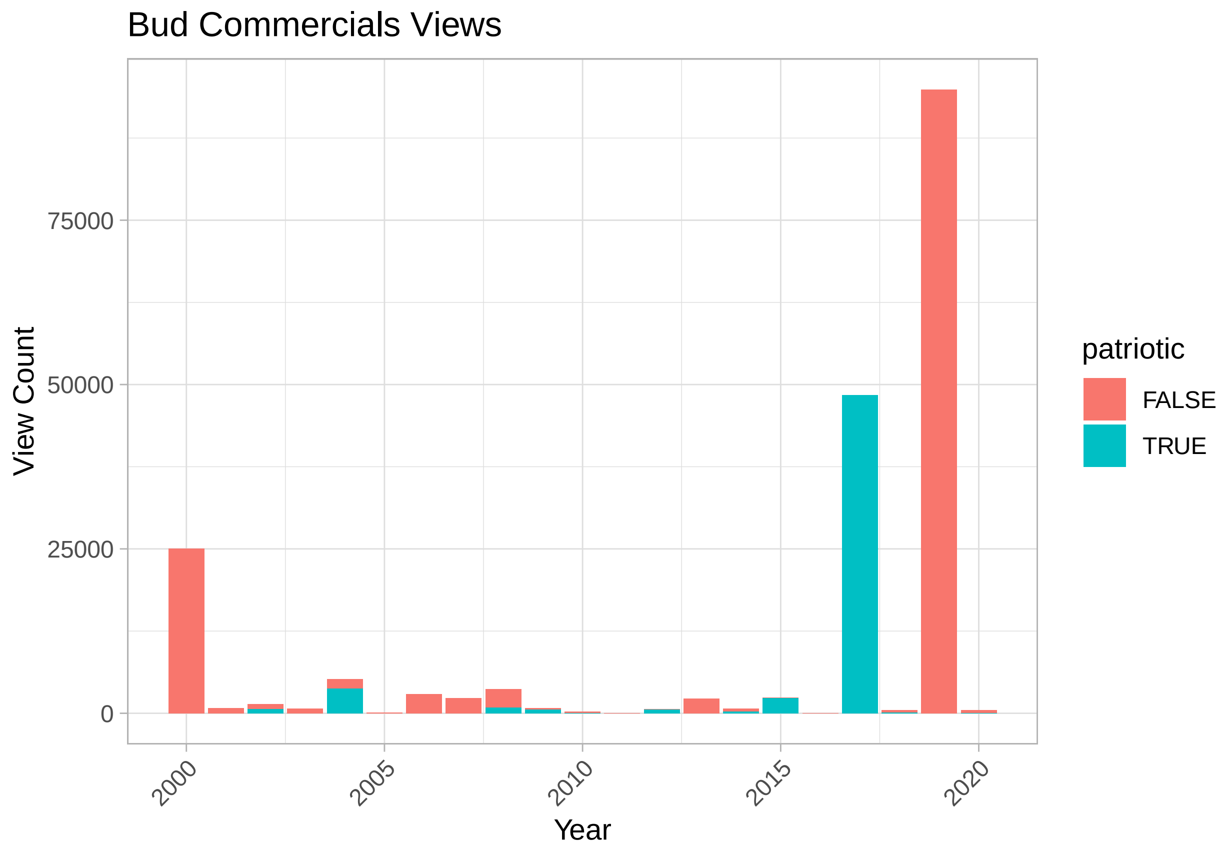
Looks like coca-cola has been less successful overall, but has more variety in their content.

```
coke <- youtube %>%
  filter(brand == "Coca-Cola")

ggplot(coke, aes(x = year, y = like_count, group = funny)) +
  geom_line(aes(color = funny)) +
  labs(title = "Coke Commercials Views",
       x = "Year",
       y = "View Count") +
  theme_light() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



One would think All-American brands like Budweiser would lean more into the patriotic side, but it seems like besides a spike in 2017 their content isn't categorized as such.

```
bud <- youtube %>%
  filter(brand == "Bud Light" | brand == "Budweiser")

ggplot(bud, aes(x = year, y = like_count, fill = patriotic)) +
  geom_bar(stat = "identity", position = "stack") +
  labs(title = "Bud Commercials Views",
       x = "Year",
       y = "View Count") +
  theme_light() +
  theme(axis.text.x = element_text(angle = 45, hjust = 1))
```



For the final chunks, I decided to run some logistic regression given the amount of binary variables in the data. Unfortunately, none of the relationships between different commercial categories and view counts were significant, but it is noteworthy that although some groups do well view-wise, that doesn't always translate to a positive relationship to likes on YouTube.

```
cols <- sapply(youtube, is.logical)
youtube[,cols] <- lapply(youtube[,cols], as.numeric)
fit <- glm(funny ~ view_count, youtube, family = "binomial")
fit_2 <-  glm(patriotic ~ view_count, youtube, family = "binomial")
fit_3 <-  glm(danger ~ view_count, youtube, family = "binomial")
fit_4 <- glm(use_sex ~ view_count, youtube, family = "binomial")

stargazer(fit, fit_2, fit_3, fit_4, title="View Count vs. Content", align=TRUE, type = "latex", digits =
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Fri, Mar 05, 2021 - 21:53:19 % Requires LaTeX packages: dcolumn

```
fit <- glm(funny ~ like_count, youtube, family = "binomial")
fit_2 <-  glm(patriotic ~ like_count, youtube, family = "binomial")
fit_3 <-  glm(danger ~ like_count, youtube, family = "binomial")
fit_4 <- glm(use_sex ~ like_count, youtube, family = "binomial")

stargazer(fit, fit_2, fit_3, fit_4, title="Like Count vs. Content", align=TRUE, type = "latex", digits =
```

% Table created by stargazer v.5.2.2 by Marek Hlavac, Harvard University. E-mail: hlavac at fas.harvard.edu % Date and time: Fri, Mar 05, 2021 - 21:53:19 % Requires LaTeX packages: dcolumn

Table 1: View Count vs. Content

	Dependent variable:			
	funny	patriotic	danger	use_sex
	(1)	(2)	(3)	(4)
view_count	0.0000000	0.0000000	−0.0000000	−0.0000002
	(0.0000000)	(0.0000000)	(0.0000000)	(0.0000002)
Constant	0.7904791***	−1.5667310***	−0.8071028***	−0.9604593**
	(0.1430179)	(0.1751745)	(0.1435651)	(0.1582249)
Observations	231	231	231	231
Log Likelihood	-143.3173000	-106.4458000	-142.4705000	-130.3355000
Akaike Inf. Crit.	290.6345000	216.8915000	288.9410000	264.6710000
Note:				*p<0.1; **p<0.05; ***p<0.0

Table 2: Like Count vs. Content

	Dependent variable:			
	funny	patriotic	danger	use_sex
	(1)	(2)	(3)	(4)
like_count	−0.0000029	0.0000052	0.0000012	−0.0001191
	(0.0000057)	(0.0000059)	(0.0000058)	(0.0001069)
Constant	0.8286775***	−1.7185440***	−0.8001970***	−0.9220614**
	(0.1469765)	(0.1878197)	(0.1462402)	(0.1611864)
Observations	225	225	225	225
Log Likelihood	-138.5608000	-96.9135300	-139.4749000	-126.5188000
Akaike Inf. Crit.	281.1217000	197.8271000	282.9498000	257.0376000
Note:				*p<0.1; **p<0.05; ***p<0.0