Raffay Plot 2

```
library(tidyverse)
```

```
## -- Attaching packages --------------------------------------- tidyverse 1.3.1 --
```

```
## v ggplot2 3.3.5     v purrr   0.3.4
## v tibble  3.1.4     v dplyr   1.0.7
## v tidyr   1.1.3     v stringr 1.4.0
## v readr   2.0.1     v forcats 0.5.1
```

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

```
library(wordcloud)
```

```
## Loading required package: RColorBrewer
```

```
library(RColorBrewer)
library(tm)
```

```
## Loading required package: NLP
```

```
##
## Attaching package: 'NLP'
```

```
## The following object is masked from 'package:ggplot2':
##
##     annotate
```

```
library(wordcloud2)
library(ggwordcloud)
library(tidytext)
library(syuzhet)
library(stringr)
library(textdata)
library(scales)
```

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

```
## The following object is masked from 'package:syuzhet':
##
##     rescale
```

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

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

```
knitr::opts_chunk$set(echo = TRUE)
youtube <- readr::read_csv("https://raw.githubusercontent.com/rfordatascience/tidytuesday/master/data/2
```

Introduction

Our second question aims to explore the difference, if any, between ads aired during election years compared to those aired during non-election years. Particularly, we want to analyze if there is any noticeable difference in the description and content between the election and non-election year ads. Answering this question involves looking at the `title` variable to analyze how these ads were described, and then analyzing the boolean content variables of `use_sex`, `patriotic`,`funny`,`celebrity`,`danger`, and `animals` to see if they had any noticeable differences in their content. Finally, the `year` variable is also needed to distinguish between election and non-election years.

We are interested in exploring this question because election years mark a significant cultural moment in the US. Therefore, we want to see whether this focus on politics translates into any noticeable effect on super bowl ads.

Approach

To analyze the description aspect of our question, we decided to use a word cloud visualization since we felt it was the most informative way to visualize what major descriptors are used for the ads. The alternative way of analyzing `title` that we considered included a bar char for top 10-20 words. However, we decided to opt for the word cloud since it provided more information in terms of relative occurrences of all the words being used (by size). Creating the word cloud involved cleaning the text (for e.g turning all words lower, removing punctuation, removing numbers) which was done by the `tm` library. We also performed sentiment analysis on the words using the `get_sentiment()` function and colored the word cloud based on the words' sentiment score.

To analyze the content aspect, we decided to opt for a column graph with percentage values of each content category as labels of the visualization. Possible alternative that we considered was a pie chart but since our column graph shows the percentage values as well as count values, we decided to go for the visualziation that maximized information.

```
# Election years in the dataset
election_year <- c(2000, 2004, 2008, 2012, 2016, 2020)
# Adding an election year variable
youtube <- youtube %>%
  mutate(election_years = ifelse(year %in% election_year, 1, 0))

test <- youtube %>%
  filter(election_years == 1) %>%
  select(title)

docs <- VCorpus(VectorSource(test))
docs <- docs %>%
  tm_map(removeNumbers) %>%
  tm_map(removePunctuation) %>%
  tm_map(stripWhitespace)
docs <- tm_map(docs, content_transformer(tolower))
docs <- tm_map(docs, removeWords, stopwords("english"))
docs <- tm_map(docs, removeWords, c(
  "super", "bowl", "commercial", "superbowl",
  "bud", "light", "budweiser", "pepsi",
  "hyundai", "doritos", "coke", "cocacola",
  "cola", "coca", "kia", "toyota"
```

```
))
dtm <- TermDocumentMatrix(docs)
matrix <- as.matrix(dtm)
words <- sort(rowSums(matrix), decreasing = TRUE)
election_df <- data.frame(word = names(words), freq = words)


election_df <- election_df %>%
  mutate(angle = sample(-45:45, nrow(election_df), replace = TRUE)) %>%
   # Package to get sentiment scores for each word in the dataset
  mutate(sentiment = get_sentiment(word,"syuzhet"))
```

```
election_df %>%
  ggplot(aes(
    label = word,
    color = sentiment,
    size = freq,
    angle = angle
  )) +
  geom_text_wordcloud() +
  scale_radius(range = c(2, 15)) +
  theme_minimal() +
  scale_color_gradient(low = "#FFD662FF", high = "#00539CFF") +
  labs(
    title = "Election Year Ad Title Word Cloud",
    subtitle = "Colored by Sentiment
      Bluer for postive sentiments
      Yellower for negative sentiment"
  )
```

```
## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'choir"' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'choir"' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'choir"' in 'mbcsToSbcs': dot substituted for <9d>

## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)): font metrics
## unknown for Unicode character U+201d


## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)): font metrics
## unknown for Unicode character U+201d

## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)): conversion
## failure on 'choir"' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)): conversion
## failure on 'choir"' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)): conversion
## failure on 'choir"' in 'mbcsToSbcs': dot substituted for <9d>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'choir"' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
```

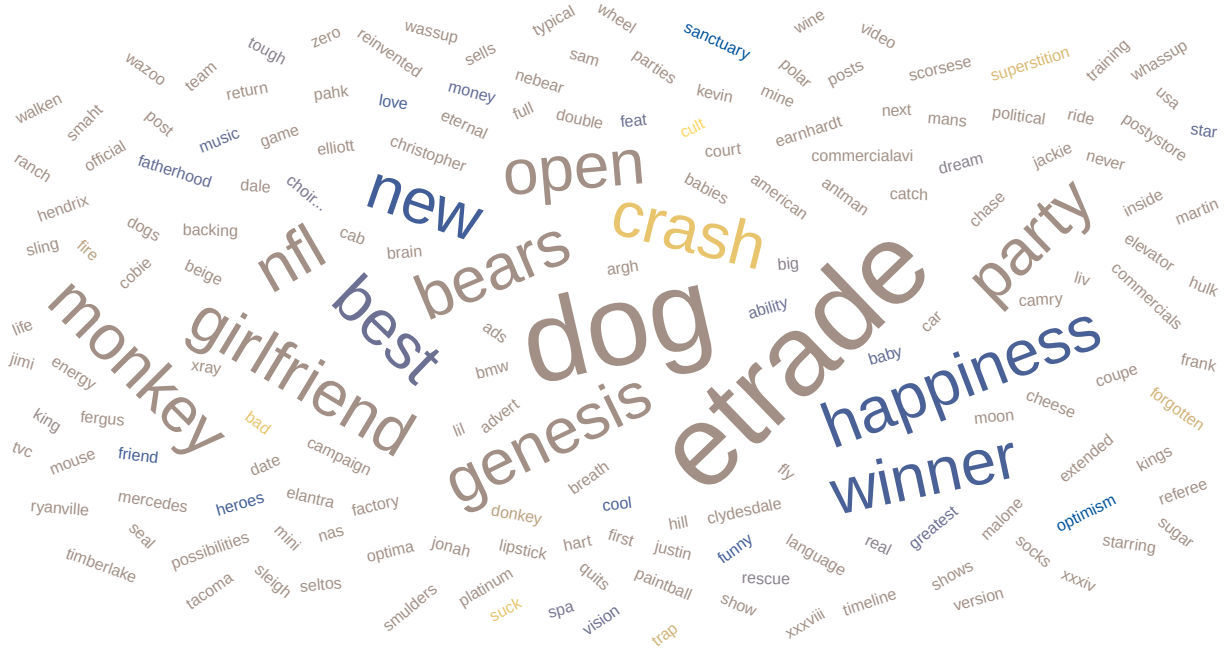
```
## conversion failure on 'choir"' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_textBounds, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'choir"' in 'mbcsToSbcs': dot substituted for <9d>

## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)): font metrics
## unknown for Unicode character U+201d

## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)): font metrics
## unknown for Unicode character U+201d

## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)): conversion
## failure on 'choir"' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)): conversion
## failure on 'choir"' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call(C_stringMetric, as.graphicsAnnot(x$label)): conversion
## failure on 'choir"' in 'mbcsToSbcs': dot substituted for <9d>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'choir"' in 'mbcsToSbcs': dot substituted for <e2>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'choir"' in 'mbcsToSbcs': dot substituted for <80>

## Warning in grid.Call.graphics(C_text, as.graphicsAnnot(x$label), x$x, x$y, :
## conversion failure on 'choir"' in 'mbcsToSbcs': dot substituted for <9d>
```

Election Year Ad Title Word Cloud

Colored by Sentiment

Bluer for postive sentiments

Yellower for negative sentiment




```
# Isolating the title variable
no_election_title <- youtube %>%
```

```
  filter(election_years == 0) %>%
  select(title)

# Data Wrangling steps
docs <- VCorpus(VectorSource(no_election_title))
docs <- docs %>%
  # Removing numbers
  tm_map(removeNumbers) %>%
  # Removing punctuation
  tm_map(removePunctuation) %>%
  # Removing whitespace
  tm_map(stripWhitespace)

# transforming words to lowercase
docs <- tm_map(docs, content_transformer(tolower))
# removing words that contain brand names
docs <- tm_map(docs, removeWords, c(
  "super", "bowl", "commercial", "superbowl",
  "bud", "light", "budweiser", "pepsi",
  "hyundai", "doritos", "coke", "cocacola",
  "cola", "coca", "the", "kia","toyota"
))

# Creating df that can be used by ggwordcloud
dtm <- TermDocumentMatrix(docs)
matrix <- as.matrix(dtm)
words <- sort(rowSums(matrix), decreasing = TRUE)
non_election_df <- data.frame(word = names(words), freq = words)

# Adding the sentiment score variable
non_election_df <- non_election_df %>%
  mutate(angle = sample(-45:45, nrow(non_election_df), replace = TRUE)) %>%
  filter(freq >= 2) %>%
  mutate(sentiment = get_sentiment(word,"syuzhet"))
```

```
non_election_df %>%
  ggplot(aes(
    label = word,
    color = sentiment,
    size = freq,
    angle = angle
  )) +
  geom_text_wordcloud() +
  scale_radius(range = c(3.5, 15)) +
  theme_minimal() +
  scale_color_gradient(low = "#FFD662FF", high = "#00539CFF")+
  labs(
    title = "Non-election Year Ad Title Word Cloud",
    subtitle = "Colored by Sentiment
      Bluer for postive sentiments
      Yellower for negative sentiment")
```

Non−election Year Ad Title Word Cloud

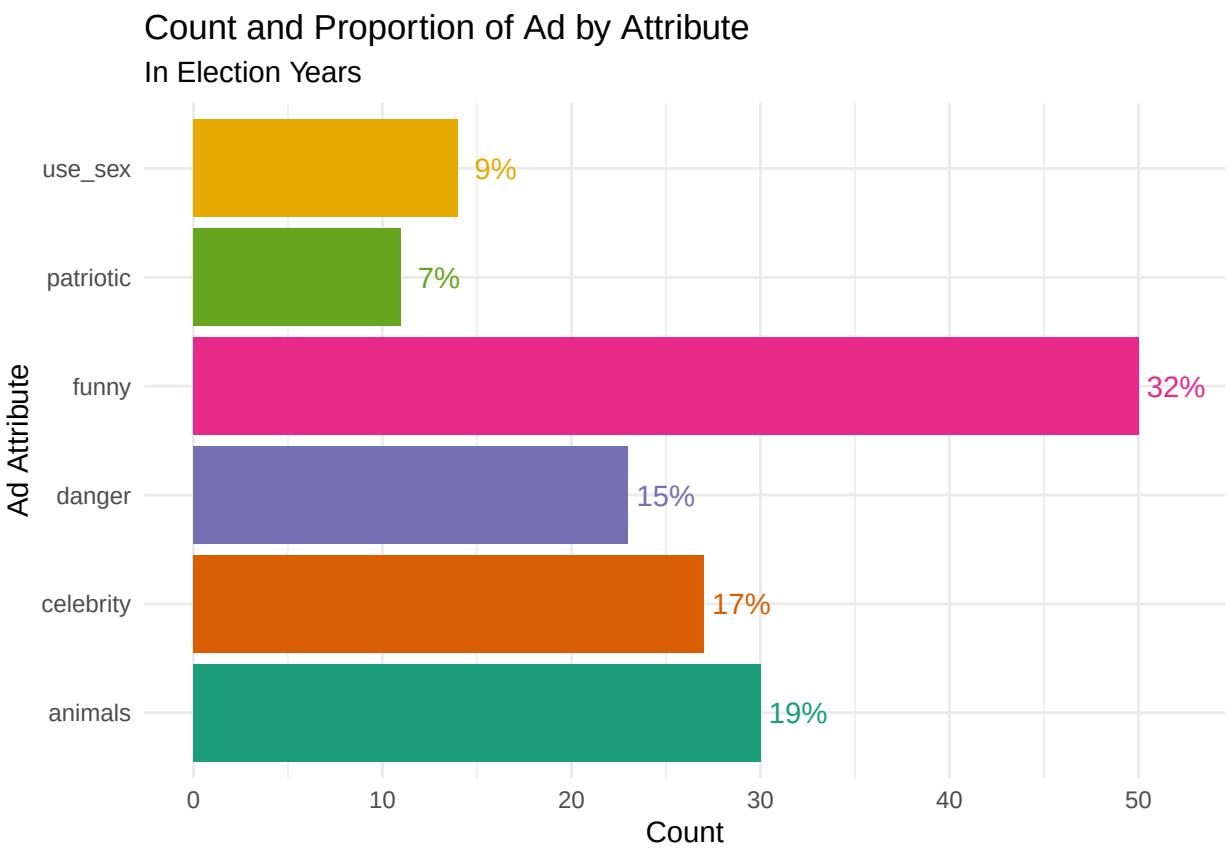
Colored by Sentiment

Bluer for postive sentiments

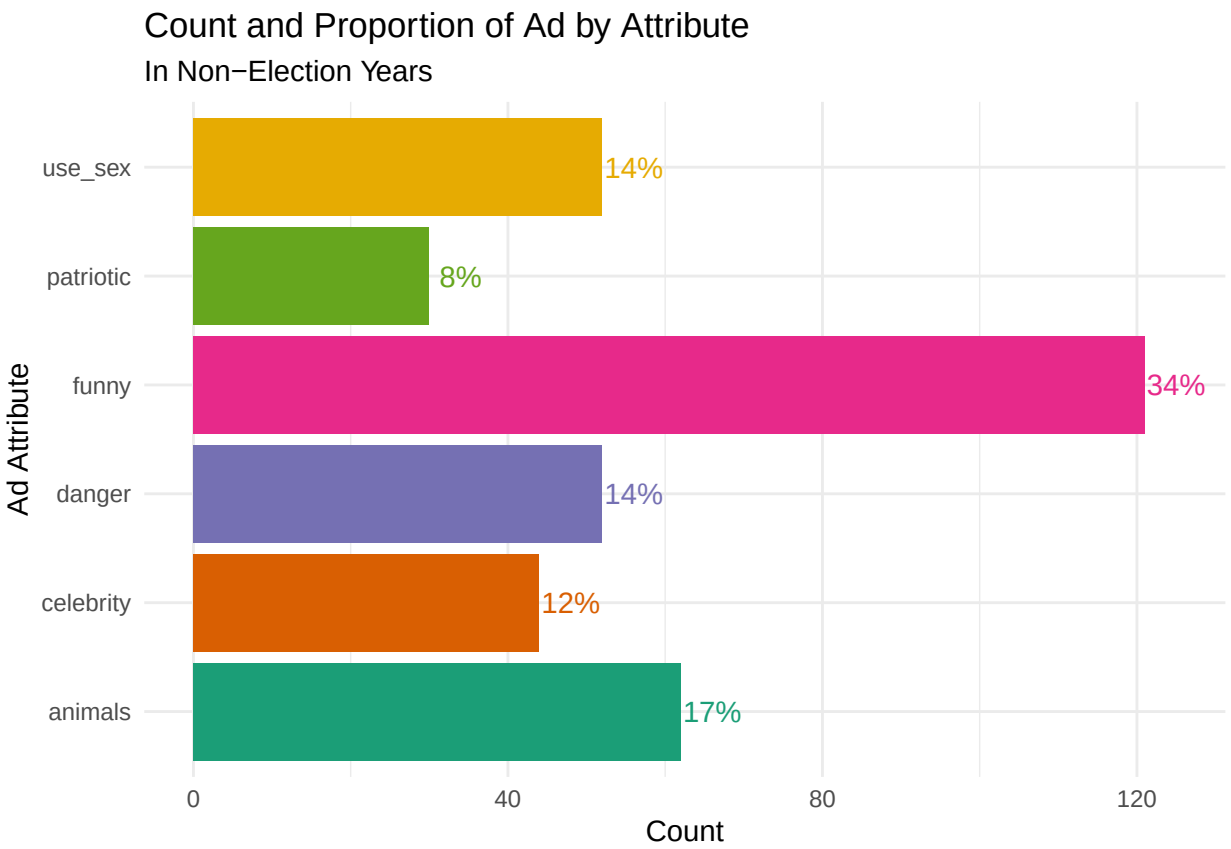
Yellower for negative sentiment

Based on the two word clouds above,

```
# Creating variable with election, using pivot_longer to get attributes in the
# same column, and creating a percentage variable
election_yr <- youtube %>%
  pivot_longer(cols = c(use_sex, funny, celebrity, patriotic, danger, animals)) %>%
  filter(election_years == 1) %>%
  filter(value == TRUE) %>%
  group_by(name) %>%
  summarise(n = n()) %>%
  mutate(perc = paste(round((n / sum(n)) * 100), "%", sep = ""))


# Creating the plot
election_yr %>%
  ggplot(aes(y = name, x = n, fill = name)) +
  geom_col(show.legend = FALSE) +
  geom_text(aes(label = perc, color = name), nudge_x = 2, show.legend = FALSE) +
  labs(
    x = "Count",
    y = "Ad Attribute",
    title = "Count and Proportion of Ad by Attribute",
    subtitle = "In Election Years"
  ) +
  scale_fill_brewer(palette = "Dark2") +
  scale_color_brewer(palette = "Dark2") +
  theme_minimal()
```




```
# Creating variable with no election, using pivot_longer to get attributes in the
# same column, and creating a percentage variable
no_election_yr <- youtube %>%
  pivot_longer(cols = c(use_sex, funny, celebrity, patriotic, danger, animals)) %>%
  filter(election_years == 0) %>%
  filter(value == TRUE) %>%
  group_by(name) %>%
  summarise(n = n()) %>%
  mutate(perc = paste(round((n / sum(n)) * 100), "%", sep = ""))

# No Election year col vizualization
no_election_yr %>%
  ggplot(aes(y = name, x = n, fill = name)) +
  geom_col(show.legend = FALSE) +
  geom_text(aes(label = perc, color = name), nudge_x = 4, show.legend = FALSE) +
  labs(
    x = "Count",
    y = "Ad Attribute",
    title = "Count and Proportion of Ad by Attribute",
    subtitle = "In Non-Election Years"
  ) +
  scale_fill_brewer(palette = "Dark2") +
  scale_color_brewer(palette = "Dark2") +
  theme_minimal()
```



https://cran.r-project.org/web/packages/ggwordcloud/vignettes/ggwordcloud.html

Discussion

We can divide the discussion of our visualizations into two parts: comparison by description and comparison by content.

When comparing our two words clouds, it immediately becomes clear that there is some overlap in the words used in Superbowl Ad titles. These include, winner, new, and nfl. Similarly, there does not seem to be any major difference in the overall sentiment makeup of these words. We can gauge this by observing that both the word clouds have a few blue words (indicating positive sentiment), very few yellow words (indicating negative sentiment), and majority brown words (indicating neutral sentiment). This goes on to indicate that there is very little difference in the titles of superbowl ads in election years compared to non-election years. This result was surprising since our team was expecting the election titles to contain more patriotic words such as “America”, “Freedom”, and “Liberty” but it appears as if the ad titles are not significantly affected by election and non-election years.

We observed a similar relation when analyzing the content of election and non-election year ads. For both these categories, there seemed to be a marginal difference in the attribute makeup, indicating that even the content of ads was very similar in election and non-election years. However, it was interesting to note that `use_sex` attribute dropped from 14% in non-election years to 9% in election years. While this could be indicative of a slight variation in content, it more more likely that this change was a coincidence.