MATH 158 - Data Description and Descriptive Statistics

John King

due Thursday, Feb 8, 2022

Introduction to Data

The data from this project comes from TidyTuesday and fivethirtyeight on GitHub. This data is a collection of the top 10 companies who posted the most advertisements during the Superbowl between the years 2000 and 2020. Because of errors in data collection, seventeen of the videos had to be removed from the data. From this, we want to compare the different qualities in the video to the youtube performance after the event.

The observational unit is each individual advertisement, and there are 10 variables. The categorical variables are logical binary variables which identify as True or False.

Quantitative Data

- Year: This variable indicates the year the Superbowl advertisement aired on TV.
- View Count: This variable indicates how many youtube views the advertisement has received.
- Like Count: This variable indicates how many youtube likes the advertisement has received.
- Dislike Count: This variable indicates how many youtube dislikes the advertisement has received.
- Comment Count: This variable indicates how many youtube comments the advertisement has received.

Categorical Data

- Funny: This variable indicates whether the advertisement is intended to be funny.
- Celebrity: This variable indicates whether a celebrity is in the advertisement.
- Danger: This variable indicates whether there is danger in the advertisement.
- Animals: This variable indicates whether there are animals in the advertisement.
- Use Sex: This variable indicates whether there is use of sexuality in the advertisement.

Summary of Statistics

As mentioned above, the dataset looked at the 10 companies who prepared advertisements between 2000 and 2020. This first graph shows how many total advertisements the 10 companies collectively posted each year. It was interesting to see how consistent certain comapnies would post one or more advertisements for every year in the Superbowl.

```r
maindata <- read.csv("Data - ExportedData.csv")
my_data <- as.data.frame(maindata)
ggplot(my_data, aes(x=year)) + geom_bar()
```

The second graph will show how many advertisements each company created for the span of Superbowls. It would be expected that companies selling beer, snacks, or cars would be the main leaders of this study. It was surprising to see the right skew of how much beer advertisements dominated the other leaders.

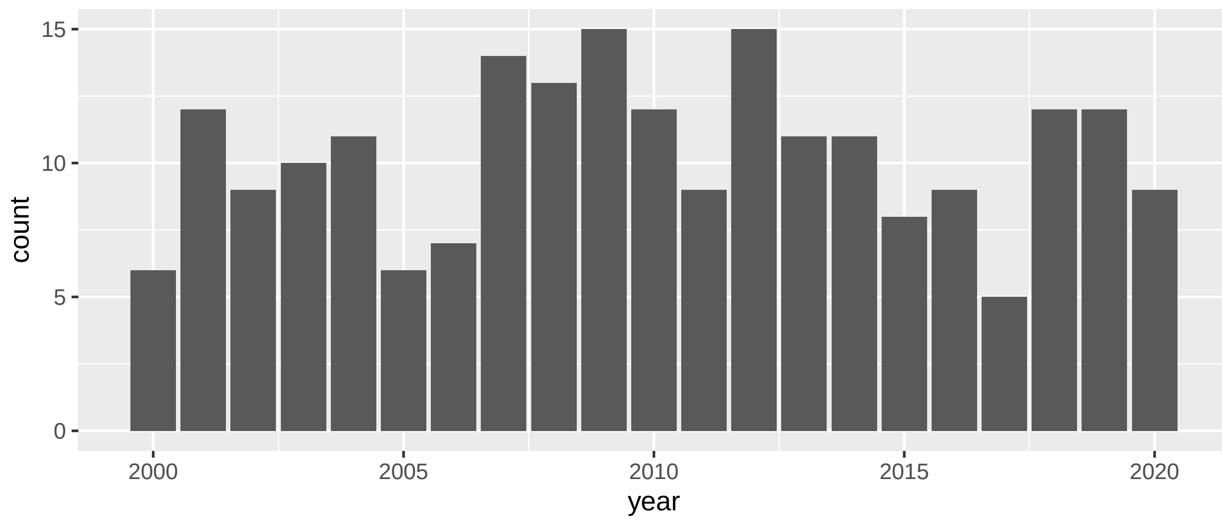

Figure 1: Advertisements by Year

```
maindata <- read.csv("Data - ExportedData.csv")
my_data <- as.data.frame(maindata)
ggplot(my_data, aes(x=brand)) + geom_bar()
```

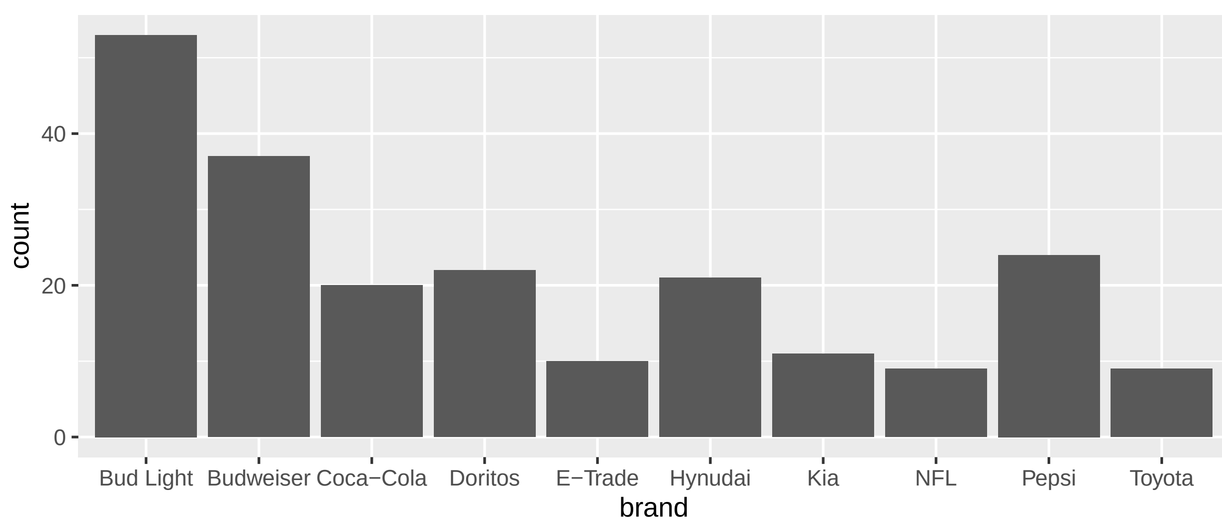

Figure 2: Advertisements by Brand

For the first table, I indicated some critical statistics encompassing all of the videos for each brand. By identifying the max column, we can see how certain brands like Doritos have very popular videos that skew their mean view count. However, Doritos, along with the NFL, still have some of the most consistent performing videos with much higher median values than the other brands.

```
maindata <- read.csv("Data - ExportedData.csv")
my_data <- as.data.frame(maindata)
df6<-maindata %>%
  group_by(brand) %>%
  summarise(Min = min(view_count),
```

```
            Max=max(view_count),
            mean = round(mean(view_count)),
            median = round(median(view_count)))
df6
```

```
## # A tibble: 10 x 5
##    brand       Min       Max    mean median
##    <fct>     <int>     <int>   <dbl>  <dbl>
##  1 Bud Light   125   7658201  262701  34565
##  2 Budweiser    10  28785122 1026244  50088
##  3 Coca-Cola   179  22849816 1618888  72245
##  4 Doritos    2985 176373378 8875610 225794
##  5 E-Trade      21   1046640  172402  50811
##  6 Hynudai      56    373684   44565   5049
##  7 Kia         518     87687   30657  17892
##  8 NFL       18670  26727063 4097798 403641
##  9 Pepsi       111    669906  121456  49830
## 10 Toyota     4873    353513  106807  32091
```

For Figure 3, I want to reveal the number of observations which identify as True or False, giving insight into how often each tactic was used to attract customers to their products. It was expected to see funny advertisements have such a high percentage, but the other categories were all consistent with each other's distribution.

```
exset <- data.frame(
  category=c("funny","celebrity","danger","animals","use_sex") ,
  percent_true=c(0.792, 0.329, 0.347, 0.426, 0.306 ))
ggplot(exset, aes(x=category, y=percent_true)) +
  geom_bar(stat = "identity")
```

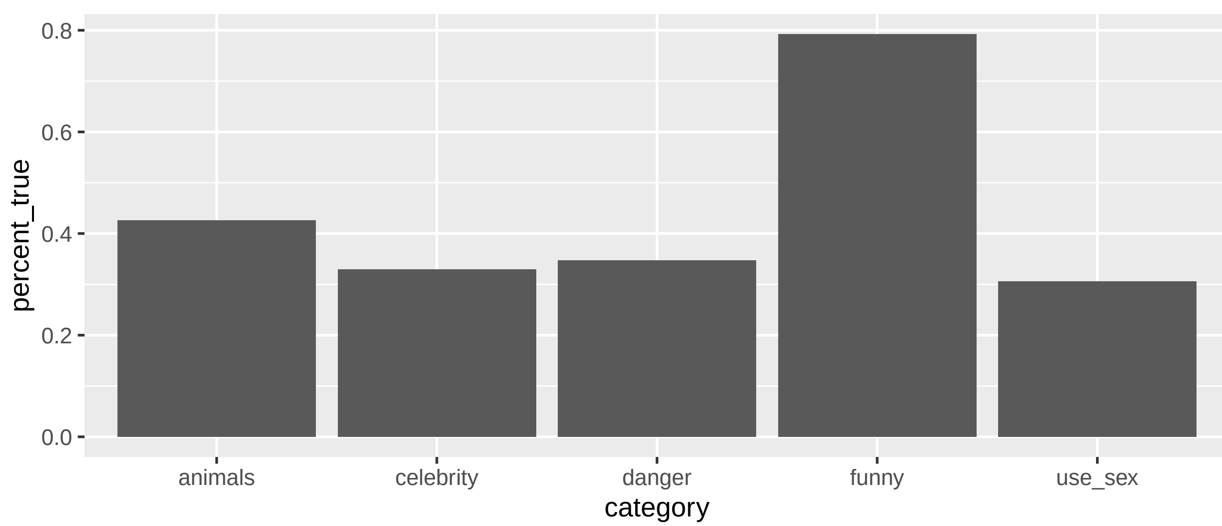

Figure 3: Percent True by Category

In this last table, I highlighted the mean and median number of views for advertisements based on responding True or False to each category. The mean values were skewed by an outlier which followed the characteristics with the highest number of views. However, the median values indicate that there could be lower views when showing celebrities or using sexuality in the advertisements.

```r
df7<-maindata %>% group_by(funny) %>% summarise(mean = round(mean(view_count)),
median = round(median(view_count)))
df8<-maindata %>% group_by(celebrity) %>% summarise(mean = round(mean(view_count)),
median = round(median(view_count)))
df9<-maindata %>% group_by(danger) %>% summarise(mean = round(mean(view_count)),
median = round(median(view_count)))
df10<-maindata %>% group_by(animals) %>% summarise(mean = round(mean(view_count)),
median = round(median(view_count)))
df11<-maindata %>% group_by(use_sex) %>% summarise(mean = round(mean(view_count)),
median = round(median(view_count)))
df_1 <- merge(df7, df8,  by = "row.names", all = TRUE)
df_2 <- merge(df_1, df9,  by = "row.names", all = TRUE)
df_3 <- merge(df_2, df10,  by = "row.names", all = TRUE)
df_4 <- merge(df_3, df11,  by = "row.names", all = TRUE)
df_5 <- df_4[ -c(1:4) ]
df_5
```

```
##   funny  mean.x median.x celebrity  mean.y median.y danger mean.x.1 median.x.1
## 1 FALSE 1359659    40358     FALSE 1768268    48978  FALSE  1680690      39814
## 2  TRUE 1559676    48546      TRUE  822187    41323   TRUE  1096275      61656
##   animals mean.y.1 median.y.1 use_sex    mean median
## 1   FALSE  1941964      41379   FALSE 1938213  48035
## 2    TRUE   692933      50850    TRUE  204302  36832
```

Final Comments

For my data, I expected some of the results I noticed when doing my analysis. However, there were a lot of surprising findings as well. The variety of variables provide interesting insight into what factors about a video corresponded to its popularity, or performance on youtube. There were very general trends of the videos with more views getting more likes, dislikes, and comments, but the amount of each in relation to one another showed unexpected results. For example, the amount of comments much more closely followed the value of dislikes than the value of likes. Also, seeing the types of advertisements change from year to year showed fascinating trends where most years would have nearly all videos within a certain category. For example, nearly every advertisement was listed as funny until 2006, and very few advertisements after 2016 use sexuality in their videos.

Even though this data targets the top 10 most frequent advertisers in the past 20 years, I think this would fairly represent the population of all Superbowl ads. Other companies would be using similar tactics, such as the categories in this study, to not only attract the viewers on the tv but also produce future popularity through other sources. Therefore, this data provides a credible insight into the characteristics and popularity of Superbowl advertisements.