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. Out of these 10 companies, this is how many total advertisements they posted each year.

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

This next graph will show how many advertisements each company created for the span of Superbowls.

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

For this next graph, I indicated some critical statistics encompassing all of the videos for each brand.

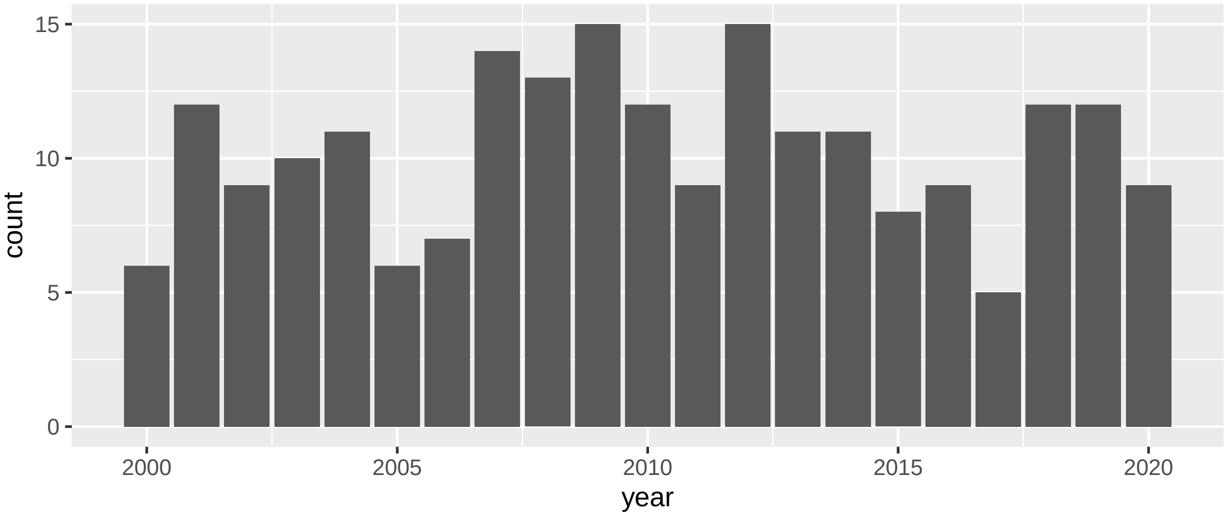

Figure 1: Advertisements by Year

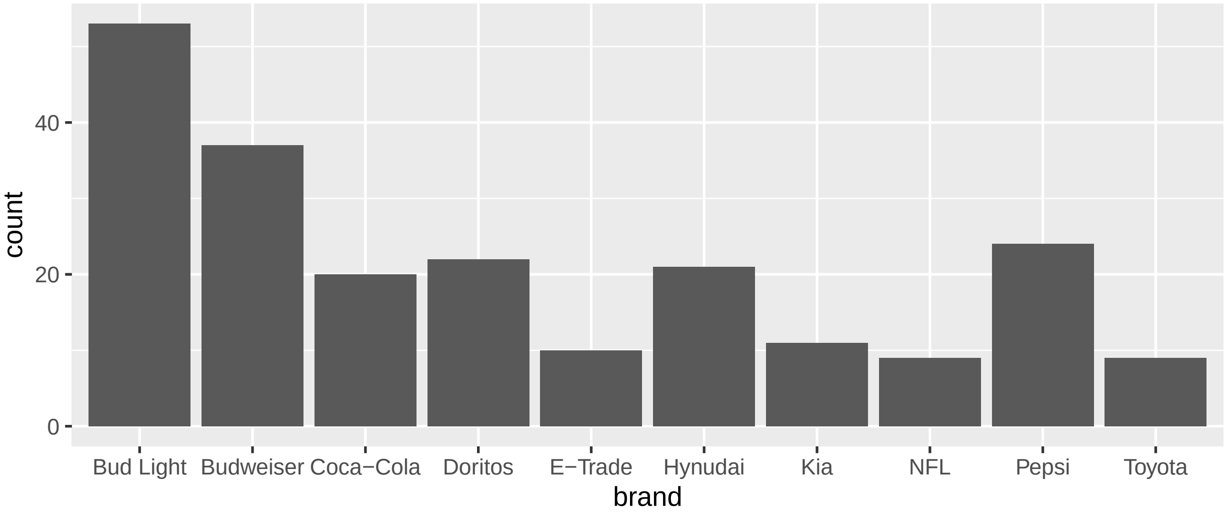

Figure 2: Advertisements by Brand

```
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 this table, I want to reveal the number of observations which identify as True or False for every categorical variable.

Object 4:

	Funny	Celebrity	Danger	Animals	Use_Sex
True	171	71	75	92	66
False	45	145	141	124	150

Next, I highlighted the mean and median number of views for advertisements based on responding True or False to each category.

Object 5:

```
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
```