Super Bowl Project

Mark Kachai

### R Markdown

This is an R Markdown document. Markdown is a simple formatting syntax for authoring HTML, PDF, and MS Word documents. For more details on using R Markdown see <http://rmarkdown.rstudio.com>.

When you click the **Knit** button a document will be generated that includes both content as well as the output of any embedded R code chunks within the document. You can embed an R code chunk like this:

Use control+Enter to run the code chunks on PC. Use command+Enter to run the code chunks on MAC.

### Load Packages

In this section, we install and load the necessary packages.

### Import Data

In this section, we import the necessary data for this lab.

# Super Bowl Case Study

You are hired as a Business Analyst to help the manager to find out if spending millions of dollars on Super Bowl ads creates social media buzz.

As a first step, we carry out the analysis using DPLYR, using superbowl.csv data.

### Dataset Exploration

# Use five functions dim(), colnames(), str(), head() and tail() to explore all the superbowl table  
# observe different columns of the table and values in the rows  
# type your answer below this line  
  
dim(superbowl)

## [1] 45 12

colnames(superbowl)

## [1] "id" "time" "new\_brand" "week\_of" "adspend" "month"   
## [7] "year" "volume" "pos" "neg" "mixed" "superbowl"

str(superbowl)

## 'data.frame': 45 obs. of 12 variables:  
## $ id : int 1 1 1 1 1 1 1 1 1 1 ...  
## $ time : int 1 2 3 4 5 6 7 8 9 10 ...  
## $ new\_brand: chr "Beetle" "Beetle" "Beetle" "Beetle" ...  
## $ week\_of : chr "2-Jan-12" "9-Jan-12" "16-Jan-12" "23-Jan-12" ...  
## $ adspend : num 2.7 4 1.3 0.7 7105 ...  
## $ month : int 1 1 1 1 1 2 2 2 2 3 ...  
## $ year : int 2012 2012 2012 2012 2012 2012 2012 2012 2012 2012 ...  
## $ volume : int 2661 3620 4138 3255 5144 8021 3728 4396 4297 4326 ...  
## $ pos : int 533 677 753 674 1199 2150 881 1021 865 895 ...  
## $ neg : int 62 87 98 99 172 435 145 172 158 148 ...  
## $ mixed : int 2066 2856 3287 2482 3773 5436 2702 3203 3274 3283 ...  
## $ superbowl: int 0 0 0 0 1 1 1 1 1 1 ...

head(superbowl)

## id time new\_brand week\_of adspend month year volume pos neg mixed  
## 1 1 1 Beetle 2-Jan-12 2.7 1 2012 2661 533 62 2066  
## 2 1 2 Beetle 9-Jan-12 4.0 1 2012 3620 677 87 2856  
## 3 1 3 Beetle 16-Jan-12 1.3 1 2012 4138 753 98 3287  
## 4 1 4 Beetle 23-Jan-12 0.7 1 2012 3255 674 99 2482  
## 5 1 5 Beetle 30-Jan-12 7105.0 1 2012 5144 1199 172 3773  
## 6 1 6 Beetle 6-Feb-12 2576.2 2 2012 8021 2150 435 5436  
## superbowl  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 1  
## 6 1

tail(superbowl)

## id time new\_brand week\_of adspend month year volume pos neg mixed  
## 40 3 10 Camaro 5-Mar-12 202.5 3 2012 108438 26236 5582 76620  
## 41 3 11 Camaro 12-Mar-12 132.3 3 2012 104014 26190 5010 72814  
## 42 3 12 Camaro 19-Mar-12 105.5 3 2012 92904 23048 4944 64912  
## 43 3 13 Camaro 26-Mar-12 534.0 3 2012 107044 26418 5742 74884  
## 44 3 14 Camaro 2-Apr-12 165.3 4 2012 85274 19522 4000 61752  
## 45 3 15 Camaro 9-Apr-12 82.4 4 2012 30078 5732 924 23422  
## superbowl  
## 40 1  
## 41 1  
## 42 1  
## 43 1  
## 44 1  
## 45 1

# tell R, new\_brand, month and superbowl columns are factor   
  
superbowl$new\_brand <- as.factor(superbowl$new\_brand)  
superbowl$month <- as.factor(superbowl$month)  
superbowl$superbowl <- as.factor(superbowl$superbowl)  
  
# tell R, week\_of is date of format "%d-%b-%y"  
  
superbowl$week\_of <- as.Date(superbowl$week\_of, "%d-%b-%y")  
  
# you can try and experiment with different date formats here:   
# https://campus.datacamp.com/courses/intermediate-r-for-finance/dates?ex=6

### Five Functions of DPLYR

# Query 1. Filter all rows where new\_brand == 'Beetle'  
  
superbowl %>% filter(new\_brand == "Beetle")

## id time new\_brand week\_of adspend month year volume pos neg mixed  
## 1 1 1 Beetle 2012-01-02 2.7 1 2012 2661 533 62 2066  
## 2 1 2 Beetle 2012-01-09 4.0 1 2012 3620 677 87 2856  
## 3 1 3 Beetle 2012-01-16 1.3 1 2012 4138 753 98 3287  
## 4 1 4 Beetle 2012-01-23 0.7 1 2012 3255 674 99 2482  
## 5 1 5 Beetle 2012-01-30 7105.0 1 2012 5144 1199 172 3773  
## 6 1 6 Beetle 2012-02-06 2576.2 2 2012 8021 2150 435 5436  
## 7 1 7 Beetle 2012-02-13 41.2 2 2012 3728 881 145 2702  
## 8 1 8 Beetle 2012-02-20 42.8 2 2012 4396 1021 172 3203  
## 9 1 9 Beetle 2012-02-27 19.8 2 2012 4297 865 158 3274  
## 10 1 10 Beetle 2012-03-05 0.3 3 2012 4326 895 148 3283  
## 11 1 11 Beetle 2012-03-12 1.4 3 2012 3870 873 113 2884  
## 12 1 12 Beetle 2012-03-19 2.5 3 2012 3840 956 135 2749  
## 13 1 13 Beetle 2012-03-26 4.3 3 2012 4240 878 109 3253  
## 14 1 14 Beetle 2012-04-02 2.3 4 2012 3760 589 84 3087  
## 15 1 15 Beetle 2012-04-09 0.9 4 2012 1885 311 44 1530  
## superbowl  
## 1 0  
## 2 0  
## 3 0  
## 4 0  
## 5 1  
## 6 1  
## 7 1  
## 8 1  
## 9 1  
## 10 1  
## 11 1  
## 12 1  
## 13 1  
## 14 1  
## 15 1

##### This query filters the superbowl table to only show those rows (observations) containing “Beetle” from the new\_brand column (variable). Based on the output from this query, we can see that there are 15 rows containing “Beetle” in the new\_brand column. From this data, we see that the highest ad spend was the week of the Superbowl as well as the week after, with ad spends of $7,105,00 and $2,576,200, respectively. These two weeks also had the largest volume of buzz metrics.

##### Another noticeable observation is that the week that had the 3rd largest volume of buzz metrics/mentions at 4,396, only had $42,800 spent on ads. During the week of the Superbowl, the volume of mentions was 5144, only 748 higher - While the amount spent on ads was about 166 times that of the amount spent on ads during the previously mentioned week, where only $42,800 was spent on ads. With that said, much less was spent on ads during the week of 2012-02-20, however, it was only 748 mentions behind the volume of mentions during the week of the Superbowl itself. This could be due to how close this date was to after the Superbowl, or it could mean that less should be spent on ads during the week of the Superbowl, given that the volume of mentions is not significantly higher than other weeks, where ad spend was significantly lower (The week of 2012-01-16, which is about two weeks before the Superbowl, also has a low ad spend but relatively high volume of buzz metrics when compared to the actual week of the Superbowl).

# Query 2. Select all the columns related to buzz metrics (volume, positive and negative mentions)  
  
superbowl %>% select(volume, pos, neg)

## volume pos neg  
## 1 2661 533 62  
## 2 3620 677 87  
## 3 4138 753 98  
## 4 3255 674 99  
## 5 5144 1199 172  
## 6 8021 2150 435  
## 7 3728 881 145  
## 8 4396 1021 172  
## 9 4297 865 158  
## 10 4326 895 148  
## 11 3870 873 113  
## 12 3840 956 135  
## 13 4240 878 109  
## 14 3760 589 84  
## 15 1885 311 44  
## 16 1014 246 54  
## 17 970 232 42  
## 18 1554 348 110  
## 19 1148 246 36  
## 20 3342 446 72  
## 21 2038 580 72  
## 22 1534 264 56  
## 23 1324 494 42  
## 24 1434 432 66  
## 25 1268 302 36  
## 26 1228 282 34  
## 27 1230 298 76  
## 28 1396 282 50  
## 29 988 262 40  
## 30 566 88 4  
## 31 78030 18528 4036  
## 32 84868 19966 4348  
## 33 91640 23230 4698  
## 34 95844 24004 4752  
## 35 98910 25140 4908  
## 36 100034 27568 5022  
## 37 94030 24234 4640  
## 38 101664 26762 5454  
## 39 107356 27746 5242  
## 40 108438 26236 5582  
## 41 104014 26190 5010  
## 42 92904 23048 4944  
## 43 107044 26418 5742  
## 44 85274 19522 4000  
## 45 30078 5732 924

##### This query selects all columns related to buzz metrics, which includes volume, pos, and neg. Based on the output from this query, we can see that for all 45 rows, the positive mentions outweigh the negative mentions. There is also a mixed column, which is not included in this output. Overall, the three brands received very positive mentions.

##### To better understand the ratio of positive to negative mentions, I created a new column in the following query, named percent\_positive.

superbowl %>% mutate(percent\_positive = pos/(pos+neg)) %>% select(new\_brand, volume, pos, neg, percent\_positive)

## new\_brand volume pos neg percent\_positive  
## 1 Beetle 2661 533 62 0.8957983  
## 2 Beetle 3620 677 87 0.8861257  
## 3 Beetle 4138 753 98 0.8848414  
## 4 Beetle 3255 674 99 0.8719276  
## 5 Beetle 5144 1199 172 0.8745441  
## 6 Beetle 8021 2150 435 0.8317215  
## 7 Beetle 3728 881 145 0.8586745  
## 8 Beetle 4396 1021 172 0.8558256  
## 9 Beetle 4297 865 158 0.8455523  
## 10 Beetle 4326 895 148 0.8581016  
## 11 Beetle 3870 873 113 0.8853955  
## 12 Beetle 3840 956 135 0.8762603  
## 13 Beetle 4240 878 109 0.8895643  
## 14 Beetle 3760 589 84 0.8751857  
## 15 Beetle 1885 311 44 0.8760563  
## 16 CR-Z 1014 246 54 0.8200000  
## 17 CR-Z 970 232 42 0.8467153  
## 18 CR-Z 1554 348 110 0.7598253  
## 19 CR-Z 1148 246 36 0.8723404  
## 20 CR-Z 3342 446 72 0.8610039  
## 21 CR-Z 2038 580 72 0.8895706  
## 22 CR-Z 1534 264 56 0.8250000  
## 23 CR-Z 1324 494 42 0.9216418  
## 24 CR-Z 1434 432 66 0.8674699  
## 25 CR-Z 1268 302 36 0.8934911  
## 26 CR-Z 1228 282 34 0.8924051  
## 27 CR-Z 1230 298 76 0.7967914  
## 28 CR-Z 1396 282 50 0.8493976  
## 29 CR-Z 988 262 40 0.8675497  
## 30 CR-Z 566 88 4 0.9565217  
## 31 Camaro 78030 18528 4036 0.8211310  
## 32 Camaro 84868 19966 4348 0.8211730  
## 33 Camaro 91640 23230 4698 0.8317817  
## 34 Camaro 95844 24004 4752 0.8347475  
## 35 Camaro 98910 25140 4908 0.8366613  
## 36 Camaro 100034 27568 5022 0.8459037  
## 37 Camaro 94030 24234 4640 0.8393018  
## 38 Camaro 101664 26762 5454 0.8307052  
## 39 Camaro 107356 27746 5242 0.8410937  
## 40 Camaro 108438 26236 5582 0.8245647  
## 41 Camaro 104014 26190 5010 0.8394231  
## 42 Camaro 92904 23048 4944 0.8233781  
## 43 Camaro 107044 26418 5742 0.8214552  
## 44 Camaro 85274 19522 4000 0.8299464  
## 45 Camaro 30078 5732 924 0.8611779

# Based on the results in this new column, we can see a very high percent of positive mentions in every row.

# Query 3. Arrange the adspend column in descending order and see which brand has the highest ad spend and in which week\_of  
  
superbowl %>% arrange(-adspend) %>% select(week\_of, adspend, new\_brand)

## week\_of adspend new\_brand  
## 1 2012-01-30 7105.0 Beetle  
## 2 2012-01-30 3977.6 Camaro  
## 3 2012-02-06 2576.2 Beetle  
## 4 2012-03-26 534.0 Camaro  
## 5 2012-02-06 272.1 Camaro  
## 6 2012-03-05 202.5 Camaro  
## 7 2012-04-02 165.3 Camaro  
## 8 2012-03-12 132.3 Camaro  
## 9 2012-03-19 105.5 Camaro  
## 10 2012-02-27 88.9 Camaro  
## 11 2012-01-16 86.9 Camaro  
## 12 2012-04-09 82.4 Camaro  
## 13 2012-01-02 75.3 Camaro  
## 14 2012-02-20 72.7 Camaro  
## 15 2012-01-23 61.2 Camaro  
## 16 2012-01-09 61.0 Camaro  
## 17 2012-02-13 59.9 Camaro  
## 18 2012-02-20 42.8 Beetle  
## 19 2012-02-13 41.2 Beetle  
## 20 2012-02-27 19.8 Beetle  
## 21 2012-03-26 4.3 Beetle  
## 22 2012-01-09 4.0 Beetle  
## 23 2012-01-02 2.7 Beetle  
## 24 2012-03-19 2.5 Beetle  
## 25 2012-01-23 2.4 CR-Z  
## 26 2012-04-02 2.3 Beetle  
## 27 2012-03-12 1.4 Beetle  
## 28 2012-01-16 1.3 Beetle  
## 29 2012-01-16 1.2 CR-Z  
## 30 2012-02-13 1.2 CR-Z  
## 31 2012-01-02 1.0 CR-Z  
## 32 2012-04-09 0.9 Beetle  
## 33 2012-01-30 0.9 CR-Z  
## 34 2012-02-20 0.9 CR-Z  
## 35 2012-02-27 0.8 CR-Z  
## 36 2012-03-12 0.8 CR-Z  
## 37 2012-01-23 0.7 Beetle  
## 38 2012-02-06 0.6 CR-Z  
## 39 2012-03-19 0.6 CR-Z  
## 40 2012-03-05 0.5 CR-Z  
## 41 2012-03-26 0.5 CR-Z  
## 42 2012-04-02 0.5 CR-Z  
## 43 2012-04-09 0.5 CR-Z  
## 44 2012-01-09 0.4 CR-Z  
## 45 2012-03-05 0.3 Beetle

##### The result of this query shows which brand has the highest ad spend in which week\_of. The adspend column is in descending order. From this, we can see that Beetle had its most ad spend (and the most ad spend compared to the other two brands), $7,105,000, during the week of 2012-01-30. Camaro had its largest ad spend of $3,977,600 during the week of 2012-01-30 as well. This makes it have the 2nd largest ad spend of the three brands. CR-Z had its largest ad spend of $2,400 during the week of 2012-01-23. Both Beetle and Camaro had their highest ad spend during the week of the Superbowl, while CR-Z had its highest adspend a week before it.

##### Overall, Beetle had the highest ad spend when compared to the other brands, during the week of the 2012-01-30, which was the week of the Superbowl.

# Query 4. Create a new column called sentiment which is a sum of positive (column name is pos) and negative (column name is neg) mentions,  
# and select only three columns for display, week\_of, new\_brand and sentiment  
  
superbowl %>% mutate(sentiment = pos+neg) %>% select(week\_of, new\_brand, sentiment)

## week\_of new\_brand sentiment  
## 1 2012-01-02 Beetle 595  
## 2 2012-01-09 Beetle 764  
## 3 2012-01-16 Beetle 851  
## 4 2012-01-23 Beetle 773  
## 5 2012-01-30 Beetle 1371  
## 6 2012-02-06 Beetle 2585  
## 7 2012-02-13 Beetle 1026  
## 8 2012-02-20 Beetle 1193  
## 9 2012-02-27 Beetle 1023  
## 10 2012-03-05 Beetle 1043  
## 11 2012-03-12 Beetle 986  
## 12 2012-03-19 Beetle 1091  
## 13 2012-03-26 Beetle 987  
## 14 2012-04-02 Beetle 673  
## 15 2012-04-09 Beetle 355  
## 16 2012-01-02 CR-Z 300  
## 17 2012-01-09 CR-Z 274  
## 18 2012-01-16 CR-Z 458  
## 19 2012-01-23 CR-Z 282  
## 20 2012-01-30 CR-Z 518  
## 21 2012-02-06 CR-Z 652  
## 22 2012-02-13 CR-Z 320  
## 23 2012-02-20 CR-Z 536  
## 24 2012-02-27 CR-Z 498  
## 25 2012-03-05 CR-Z 338  
## 26 2012-03-12 CR-Z 316  
## 27 2012-03-19 CR-Z 374  
## 28 2012-03-26 CR-Z 332  
## 29 2012-04-02 CR-Z 302  
## 30 2012-04-09 CR-Z 92  
## 31 2012-01-02 Camaro 22564  
## 32 2012-01-09 Camaro 24314  
## 33 2012-01-16 Camaro 27928  
## 34 2012-01-23 Camaro 28756  
## 35 2012-01-30 Camaro 30048  
## 36 2012-02-06 Camaro 32590  
## 37 2012-02-13 Camaro 28874  
## 38 2012-02-20 Camaro 32216  
## 39 2012-02-27 Camaro 32988  
## 40 2012-03-05 Camaro 31818  
## 41 2012-03-12 Camaro 31200  
## 42 2012-03-19 Camaro 27992  
## 43 2012-03-26 Camaro 32160  
## 44 2012-04-02 Camaro 23522  
## 45 2012-04-09 Camaro 6656

##### The results of this query show the sum of positive and negative mentions for each brand in a column named sentiment. From these results, we can see that Camaro has a significantly high sentiment for almost every week in the dataset. Beetle has the second largest volume of sentiment. CR-Z has the least amount of sentiment overall when compared to the other two brands. The largest volume of sentiment experienced by Camaro was 32,988 during the week of 2012-02-27 (a few weeks after the Superbowl). The largest volume of sentiment experienced by Beetle was 2,585 during the week of 2012-02-06 (during the week after the Superbowl). Lastly, the largest volume of sentiment experienced by CR-Z was 652 during the week of 2012-02-06 (during the week after the Superbowl).

##### Overall, Camaro easily had the highest volume of sentiment among all the brands. Camaro was able to maintain a significantly high volume of sentiment weeks before and after the Superbowl, meaning its ads were continuously highly effective.

### Section 1: Univariate Statistics

# First, compute the \*mean\*, \*variance\*, and \*standard deviation\* of \*\*adspend\*\* for each \*\*new\_brand\*\* before the Super Bowl and after the Super Bowl, using the relevant \*R\* functions.  
  
superbowl %>% group\_by(new\_brand, superbowl) %>% summarise(mean=mean(adspend), variance=var(adspend), sd=sd(adspend))

## `summarise()` has grouped output by 'new\_brand'. You can override using the  
## `.groups` argument.

## # A tibble: 6 × 5  
## # Groups: new\_brand [3]  
## new\_brand superbowl mean variance sd  
## <fct> <fct> <dbl> <dbl> <dbl>  
## 1 Beetle 0 2.17 2.18 1.48   
## 2 Beetle 1 891. 4839676. 2200.   
## 3 Camaro 0 71.1 156. 12.5   
## 4 Camaro 1 518. 1335443. 1156.   
## 5 CR-Z 0 1.25 0.703 0.839  
## 6 CR-Z 1 0.709 0.0529 0.230

##### The output of this query shows the mean, variance, and standard deviation of adpsend for each new brand before and after the Super Bowl.

##### For Beetle and Camaro, the mean ad spend significantly increased after (and during the week of) the Super Bowl, while the mean ad spend for CR-Z decreased. The variance for ad spend between Beetle and Camaro significantly increases from before to after the Super Bowl. Likewise, the standard deviation also significantly increases from before to after the Super Bowl for Beetle and Camaro as well.

# Second, create a new column in superbowl dataset, call it \*\*pos\_prop\*\*, and calculate the proportion of positive mentions for each \*\*new\_brand\*\* and each \*\*week\*\* (each row). Similarly, create a new column in superbowl dataset, call it \*\*neg\_prop\*\*, and calculate the proportion of negative mentions for each \*\*new\_brand\*\* and each \*\*week\*\*.  
  
superbowl <- superbowl %>% mutate(pos\_prop=pos/volume, neg\_prop=neg/volume)

##### This query added two new columns to the superbowl dataset: pos\_prop and neg\_prop. The column ‘pos\_prop’ shows the proportion of positive mentions for each new\_brand and each week (row). The column ‘neg\_prop’ shows the proportion of negative mentions for each new\_brand and each week (row).

# Third, compute the \*mean\* and \*standard deviation\* of \*\*volume\*\*, \*\*pos\_prop\*\* and \*\*neg\_prop\*\* for each \*\*new\_brand\*\* before the Super Bowl and after the Super Bowl.  
  
# The mean and standard deviation of volume for each new\_brand before the Super Bowl and after the Super Bowl.  
  
superbowl %>% group\_by(new\_brand, superbowl) %>% summarise(mean=mean(volume), sd=sd(volume))

## `summarise()` has grouped output by 'new\_brand'. You can override using the  
## `.groups` argument.

## # A tibble: 6 × 4  
## # Groups: new\_brand [3]  
## new\_brand superbowl mean sd  
## <fct> <fct> <dbl> <dbl>  
## 1 Beetle 0 3418. 622.  
## 2 Beetle 1 4319. 1463.  
## 3 Camaro 0 87596. 7817.  
## 4 Camaro 1 93613. 22222.  
## 5 CR-Z 0 1172. 266.  
## 6 CR-Z 1 1486. 712.

# The mean and standard deviation of pos\_prop for each new\_brand before the Super Bowl and after the Super Bowl.  
  
superbowl %>% group\_by(new\_brand, superbowl) %>% summarise(mean=mean(pos\_prop), sd=sd(pos\_prop))

## `summarise()` has grouped output by 'new\_brand'. You can override using the  
## `.groups` argument.

## # A tibble: 6 × 4  
## # Groups: new\_brand [3]  
## new\_brand superbowl mean sd  
## <fct> <fct> <dbl> <dbl>  
## 1 Beetle 0 0.194 0.0116   
## 2 Beetle 1 0.216 0.0337   
## 3 Camaro 0 0.244 0.00915  
## 4 Camaro 1 0.247 0.0222   
## 5 CR-Z 0 0.230 0.0133   
## 6 CR-Z 1 0.236 0.0697

# The mean and standard deviation of neg\_prop for each new\_brand before the Super Bowl and after the Super Bowl.  
  
superbowl %>% group\_by(new\_brand, superbowl) %>% summarise(mean=mean(neg\_prop), sd=sd(neg\_prop))

## `summarise()` has grouped output by 'new\_brand'. You can override using the  
## `.groups` argument.

## # A tibble: 6 × 4  
## # Groups: new\_brand [3]  
## new\_brand superbowl mean sd  
## <fct> <fct> <dbl> <dbl>  
## 1 Beetle 0 0.0254 0.00338   
## 2 Beetle 1 0.0339 0.00900   
## 3 Camaro 0 0.0510 0.000940  
## 4 Camaro 1 0.0487 0.00638   
## 5 CR-Z 0 0.0497 0.0167   
## 6 CR-Z 1 0.0339 0.0139

### Questions:

##### What brand has the **highest mean ad spend** after the Super Bowl? Has the mean ad spending increased after the Super Bowl for that brand? By how much?

superbowl %>% group\_by(new\_brand, superbowl) %>% summarise(mean=mean(adspend))

## `summarise()` has grouped output by 'new\_brand'. You can override using the  
## `.groups` argument.

## # A tibble: 6 × 3  
## # Groups: new\_brand [3]  
## new\_brand superbowl mean  
## <fct> <fct> <dbl>  
## 1 Beetle 0 2.17   
## 2 Beetle 1 891.   
## 3 Camaro 0 71.1   
## 4 Camaro 1 518.   
## 5 CR-Z 0 1.25   
## 6 CR-Z 1 0.709

##### From this query, we find the brand that has the highest mean ad spend after the Super Bowl to be Beetle, with a mean ad spend of $891,000. The mean ad spending increased significantly after the Super Bowl, from an average of $2,170 to $891,000. The average ad spend increased by $888,830 from before to after the Super Bowl.

##### What brand has the **highest mean** volume of mentions after the Super Bowl ? On average, what proportion of mentions are positive and what proportion of mentions are negative for that brand after the Super Bowl?

superbowl %>% filter(superbowl==1) %>% group\_by(new\_brand) %>% summarise(mean\_volume=mean(volume), mean\_pos\_prop=mean(pos\_prop), mean\_neg\_prop=mean(neg\_prop))

## # A tibble: 3 × 4  
## new\_brand mean\_volume mean\_pos\_prop mean\_neg\_prop  
## <fct> <dbl> <dbl> <dbl>  
## 1 Beetle 4319. 0.216 0.0339  
## 2 Camaro 93613. 0.247 0.0487  
## 3 CR-Z 1486. 0.236 0.0339

##### From this query, we find the brand that has the highest mean volume of mentions after the Super Bowl to be Camaro, with a mean volume of 93,613 mentions after the Super Bowl. On average, the proportions of mentions that are positive for Camaro after the Super Bowl is 0.247. The proportions of mentions that are negative for Camaro after the Super Bowl, on average, is 0.0487. So, on average, there is a higher proportion of positive mentions for Camaro than negative mentions after the Super Bowl.

##### What brand has the **lowest variability** in volume of mentions after the Super Bowl?

superbowl %>% filter(superbowl==1) %>% group\_by(new\_brand) %>% summarise(variability=sd(volume))

## # A tibble: 3 × 2  
## new\_brand variability  
## <fct> <dbl>  
## 1 Beetle 1463.  
## 2 Camaro 22222.  
## 3 CR-Z 712.

##### The output of this query shows the variability in volume of mentions for each brand after the Super Bowl. Based on this query, the brand that has the lowest variability in volume of mentions after the Super Bowl is CR-Z, with a standard deviation of 712.

### Section 2: Bivariate Statistics

# First, compute the \*correlation\* between \*\*adspend\*\* and \*\*volume\*\* in superbowl dataset, before the Super Bowl and after the Super Bowl, using the relevant \*R\* functions.  
  
superbowl %>% group\_by(superbowl) %>% summarise(correlation=cor(adspend, volume))

## # A tibble: 2 × 2  
## superbowl correlation  
## <fct> <dbl>  
## 1 0 0.975   
## 2 1 0.0513

##### The result of this query shows that there is an extremely strong, positive correlation between adspend and volume before the Super Bowl. On the other hand, there is almost no correlation between adspend and volume after the Super Bowl.

# Similarly, compute the \*correlation\* between \*\*adspend\*\* and \*\*pos\_prop\*\* and between \*\*adspend\*\* and \*\*neg\_prop\*\* in superbowl dataset, before the Super Bowl and after the Super Bowl.  
  
superbowl %>% group\_by(superbowl) %>% summarise(correlation=cor(adspend, pos\_prop))

## # A tibble: 2 × 2  
## superbowl correlation  
## <fct> <dbl>  
## 1 0 0.642   
## 2 1 0.0918

superbowl %>% group\_by(superbowl) %>% summarise(correlation=cor(adspend, neg\_prop))

## # A tibble: 2 × 2  
## superbowl correlation  
## <fct> <dbl>  
## 1 0 0.419  
## 2 1 0.114

##### The result of this query shows that there is a very strong, positive correlation between adspend and the proportion of positive mentions before the Super Bowl, while there is almost no correlation between adspend and pos\_prop after the Super Bowl. Also, there is a moderate, positive correlation between adspend and neg\_prop before the Super Bowl, while there is a very weak, positive correlation (almost no correlation) after the Super Bowl.

# Second, compute the \*covariance\* and \*correlation\* between \*\*adspend\*\* and \*\*volume\*\*, between \*\*adspend\*\* and \*\*pos\_prop\*\*, and between \*\*adspend\*\* and \*\*neg\_prop\*\* for each \*\*new\_brand\*\*, before the Super Bowl and after the Super Bowl.  
  
superbowl %>% group\_by(new\_brand, superbowl) %>% summarise(covariance=cov(adspend, volume), correlation=cor(adspend,volume))

## `summarise()` has grouped output by 'new\_brand'. You can override using the  
## `.groups` argument.

## # A tibble: 6 × 4  
## # Groups: new\_brand [3]  
## new\_brand superbowl covariance correlation  
## <fct> <fct> <dbl> <dbl>  
## 1 Beetle 0 -139. -0.152  
## 2 Beetle 1 1537337. 0.478  
## 3 Camaro 0 -10128. -0.104  
## 4 Camaro 1 2950874. 0.115  
## 5 CR-Z 0 54.8 0.246  
## 6 CR-Z 1 64.7 0.396

superbowl %>% group\_by(new\_brand, superbowl) %>% summarise(covariance=cov(adspend, pos\_prop), correlation=cor(adspend, pos\_prop))

## `summarise()` has grouped output by 'new\_brand'. You can override using the  
## `.groups` argument.

## # A tibble: 6 × 4  
## # Groups: new\_brand [3]  
## new\_brand superbowl covariance correlation  
## <fct> <fct> <dbl> <dbl>  
## 1 Beetle 0 -0.00606 -0.354   
## 2 Beetle 1 25.2 0.340   
## 3 Camaro 0 0.0490 0.429   
## 4 Camaro 1 3.10 0.121   
## 5 CR-Z 0 -0.00957 -0.861   
## 6 CR-Z 1 -0.000801 -0.0500

superbowl %>% group\_by(new\_brand, superbowl) %>% summarise(covariance=cov(adspend, neg\_prop), correlation=cor(adspend,neg\_prop))

## `summarise()` has grouped output by 'new\_brand'. You can override using the  
## `.groups` argument.

## # A tibble: 6 × 4  
## # Groups: new\_brand [3]  
## new\_brand superbowl covariance correlation  
## <fct> <fct> <dbl> <dbl>  
## 1 Beetle 0 -0.00316 -0.633   
## 2 Beetle 1 4.99 0.252   
## 3 Camaro 0 0.00631 0.538   
## 4 Camaro 1 0.627 0.0850  
## 5 CR-Z 0 -0.00586 -0.419   
## 6 CR-Z 1 0.0000828 0.0259

### Super Bowl Data Visualizations

##### This section contains the queries used to create different bar plots for the superbowl data using ggplot2.

#Bar plot showing the average ad spend for each new brand before and after the Super Bowl.  
  
avg\_adspend\_bar\_chart <- ggplot(superbowl, aes(x = new\_brand, y = adspend, fill = superbowl)) +  
 geom\_bar(stat = "summary", fun = "mean", position = "dodge") +  
 geom\_text(stat = "summary",  
 fun = "mean",  
 aes(label = sprintf("%.2f", after\_stat(y)), group = superbowl),  
 position = position\_dodge(width = 0.9),  
 vjust = -0.5,   
 size = 4) +  
 labs(title = "Average Ad Spend Before and After Super Bowl",  
 x = "Car Brands",  
 y = "Average Ad Spend ($1,000's)") +  
 scale\_fill\_manual(values = c("0" = "darkblue", "1" = "darkorange"), labels = c("Before Super Bowl", "After Super Bowl")) +  
 theme\_minimal()  
  
print(avg\_adspend\_bar\_chart)

A graph of a number of people

Description automatically generated with medium confidence

#Bar plot showing the average volume of mentions for each new brand before and after the Super Bowl.  
  
avg\_volume\_bar\_chart <- ggplot(superbowl, aes(x = new\_brand, y = volume, fill = superbowl)) +  
 geom\_bar(stat = "summary", fun = "mean", position = "dodge") +  
 geom\_text(stat = "summary",  
 fun = "mean",  
 aes(label = sprintf("%.0f", after\_stat(y)), group = superbowl),  
 position = position\_dodge(width = 0.9),  
 vjust = -0.5,   
 size = 4) +  
 labs(title = "Average Volume of Mentions Before and After Super Bowl",  
 x = "Car Brands",  
 y = "Average Volume",  
 caption = "#% = Percent Change in Volume of Mentions") +  
 scale\_fill\_manual(values = c("0" = "darkblue", "1" = "darkorange"), labels = c("Before Super Bowl", "After Super Bowl")) +  
 theme\_minimal() +  
 annotate(geom = "text", x = "Beetle", y = 17000, label = "+26.4%", color = "darkgreen", size = 4) +  
 annotate(geom = "text", x = "Camaro", y = 105000, label = "+6.87%", color = "darkgreen", size = 4) +  
 annotate(geom = "text", x = "CR-Z", y = 17000, label = "+26.8%", color = "darkgreen", size = 4)  
  
print(avg\_volume\_bar\_chart)

A graph of a number of people

Description automatically generated with medium confidence

### Questions:

##### What is the *correlation* between **adspend** and **volume** before AND after the Super Bowl? Does that imply a positive, negative or no relationship? Is the relationship strong?

superbowl %>% group\_by(superbowl) %>% summarise(correlation=cor(adspend, volume))

## # A tibble: 2 × 2  
## superbowl correlation  
## <fct> <dbl>  
## 1 0 0.975   
## 2 1 0.0513

##### The output of this query shows the correlation between ad spend and volume before the Super Bowl to be 0.975, while the correlation between ad spend and volume after the Super Bowl is shown to be 0.0513. The correlation between ad spend and volume before the Super Bowl, 0.975, implies a very strong, positive relationship between ad spend and volume of mentions. The correlation between ad spend and volume after the Super Bowl, 0.0513, although positive, implies almost no relationship between ad spend and volume of mentions since this value is very close to 0.

##### What is the *correlation* between **adspend** and **pos\_prop** before AND after the Super Bowl? Does that imply a positive, negative or no relationship? Is the relationship strong? What about the relationship between **adspend** and **neg\_prop** before AND after the Super Bowl?

superbowl %>% group\_by(superbowl) %>% summarise(correlation\_adspend\_and\_pos\_prop=cor(adspend, pos\_prop), correlation\_adspend\_and\_neg\_prop=cor(adspend, neg\_prop))

## # A tibble: 2 × 3  
## superbowl correlation\_adspend\_and\_pos\_prop correlation\_adspend\_and\_neg\_prop  
## <fct> <dbl> <dbl>  
## 1 0 0.642 0.419  
## 2 1 0.0918 0.114

##### Based on the output of this query, the correlation between adspend and pos\_prop before the Super Bowl is 0.642, which implies a strong, positive relationship between the two. The correlation between adspend and pos\_prop after the Super Bowl is 0.0918, which implies a very weak, yet positive relationship between the two - however, it is safer to say this value implies almost no relationship between the two since the correlation value of 0.0918 is very close to 0.

##### Also, the correlation between adspend and neg\_prop before the Super Bowl is 0.419, which implies a normal, positive relationship between the two. The correlation between adspend and neg\_prop after the Super Bowl is 0.114, which implies a that there is almost no relationship between the two since the correlation value of 0.114 is very close to 0.

##### Which brand’s ad spending did create a social media buzz (either positive or negative mentions)? Use both univariate and bivariate statistics for each brand.

# Beetle:  
# Univariate statistics:  
superbowl %>% filter(new\_brand=="Beetle") %>% group\_by(new\_brand, superbowl) %>% summarise(mean\_adspend=mean(adspend), mean\_volume\_mentions=mean(volume), mean\_pos\_prop=mean(pos\_prop), mean\_neg\_prop=mean(neg\_prop))

## `summarise()` has grouped output by 'new\_brand'. You can override using the  
## `.groups` argument.

## # A tibble: 2 × 6  
## # Groups: new\_brand [1]  
## new\_brand superbowl mean\_adspend mean\_volume\_mentions mean\_pos\_prop  
## <fct> <fct> <dbl> <dbl> <dbl>  
## 1 Beetle 0 2.17 3418. 0.194  
## 2 Beetle 1 891. 4319. 0.216  
## # ℹ 1 more variable: mean\_neg\_prop <dbl>

# Bivariate statistics:  
superbowl %>% filter(new\_brand=="Beetle") %>% group\_by(new\_brand, superbowl) %>% summarise(cor\_adspend\_and\_volume=cor(adspend,volume), cor\_adspend\_and\_pos\_prop=cor(adspend,pos\_prop), cor\_adspend\_and\_neg\_prop=cor(adspend,neg\_prop))

## `summarise()` has grouped output by 'new\_brand'. You can override using the  
## `.groups` argument.

## # A tibble: 2 × 5  
## # Groups: new\_brand [1]  
## new\_brand superbowl cor\_adspend\_and\_volume cor\_adspend\_and\_pos\_prop  
## <fct> <fct> <dbl> <dbl>  
## 1 Beetle 0 -0.152 -0.354  
## 2 Beetle 1 0.478 0.340  
## # ℹ 1 more variable: cor\_adspend\_and\_neg\_prop <dbl>

##### Based on the results of the univariate statistics, mean ad spend for Beetle increased by $888,830 from before to after the Super Bowl. This is a very substantial increase in ad spending, yet the mean volume of overall mentions only went up by less than 1,000 mentions. The mean pos\_prop went up by only 0.022, while the mean neg\_prop increased by 0.0085. Although the increase in mean neg\_prop is minuscule, this does not mean the jump in ad spending was justified, as the mean proportion of positive mentions barely increased.

##### The bivariate statistics return the correlation between adspend and volume, pos\_prop, and neg\_prop for Beetle. The correlation between ad spend and volume, pos\_prop, and neg\_prop was all negative before the Super Bowl. It should be noted that the negative correlation between ad spend and neg\_prop before the Super Bowl for Beetle, -0.633, should be considered substantial as it is close to -1.0. This strong, negative correlation between ad spend and neg\_prop before the Super Bowl for Beetle means that on average, as ad spending increased for Beetle before the Super Bowl, the proportion of negative mentions decreased. The correlation between ad spend and volume, pos\_prop, and neg\_prop was all positive after the Super Bowl, however, I would say only the correlation between ad spend and volume was moderately significant, with a positive correlation value of 0.478.

# CR-Z:  
# Univariate statistics:  
superbowl %>% filter(new\_brand=="CR-Z") %>% group\_by(new\_brand, superbowl) %>% summarise(mean\_adspend=mean(adspend), mean\_volume\_mentions=mean(volume), mean\_pos\_prop=mean(pos\_prop), mean\_neg\_prop=mean(neg\_prop))

## `summarise()` has grouped output by 'new\_brand'. You can override using the  
## `.groups` argument.

## # A tibble: 2 × 6  
## # Groups: new\_brand [1]  
## new\_brand superbowl mean\_adspend mean\_volume\_mentions mean\_pos\_prop  
## <fct> <fct> <dbl> <dbl> <dbl>  
## 1 CR-Z 0 1.25 1172. 0.230  
## 2 CR-Z 1 0.709 1486. 0.236  
## # ℹ 1 more variable: mean\_neg\_prop <dbl>

# Bivariate statistics:   
superbowl %>% filter(new\_brand=="CR-Z") %>% group\_by(new\_brand, superbowl) %>% summarise(cor\_adspend\_and\_volume=cor(adspend,volume), cor\_adspend\_and\_pos\_prop=cor(adspend,pos\_prop), cor\_adspend\_and\_neg\_prop=cor(adspend,neg\_prop))

## `summarise()` has grouped output by 'new\_brand'. You can override using the  
## `.groups` argument.

## # A tibble: 2 × 5  
## # Groups: new\_brand [1]  
## new\_brand superbowl cor\_adspend\_and\_volume cor\_adspend\_and\_pos\_prop  
## <fct> <fct> <dbl> <dbl>  
## 1 CR-Z 0 0.246 -0.861   
## 2 CR-Z 1 0.396 -0.0500  
## # ℹ 1 more variable: cor\_adspend\_and\_neg\_prop <dbl>

##### The univariate statistics return the mean ad spend for CR-Z, which can then be compared to the other columns returned in the query, like mean volume of mentions and the mean positive and negative proportion of mentions. From this query, we can see that the mean ad spend for CR-Z decreased from before to after the Super Bowl by $541. Interestingly, although ad spending was significantly reduced after the Super Bowl, the mean volume of mentions increased by 314, with a very slight increase in the average pos\_prop as well.

##### The bivariate statistics return the correlation between adspend and volume, pos\_prop, and neg\_prop for CR-Z. The correlation between ad spend and volume before and after the Super Bowl is positive, however the correlation between the two after the Super Bowl is more substantial, with a value of 0.396. This value shows a moderate, positive correlation between adspend and volume of mentions for CR-Z after the Super Bowl. Also, the correlation between ad spend and both pos and neg prop before the Super Bowl are both negative, but the strength of the correlation varies. Before the Super Bowl, there is a very strong, negative correlation between ad spend and pos\_prop for CR-Z. This implies that on average, as ad spending increased, the proportion of positive mentions went down. There is also a moderate, negative correlation between ad spend and neg\_prop for CR-Z before the Super Bowl, meaning that on average, as ad spending increased, negative mentions would slightly decrease. After the Super Bowl, there is no substantial correlation between ad spend and either pos\_prop or neg\_prop for CR-Z.

# Camaro:  
# Univariate statistics:  
superbowl %>% filter(new\_brand=="Camaro") %>% group\_by(new\_brand, superbowl) %>% summarise(mean\_adspend=mean(adspend), mean\_volume\_mentions=mean(volume), mean\_pos\_prop=mean(pos\_prop), mean\_neg\_prop=mean(neg\_prop))

## `summarise()` has grouped output by 'new\_brand'. You can override using the  
## `.groups` argument.

## # A tibble: 2 × 6  
## # Groups: new\_brand [1]  
## new\_brand superbowl mean\_adspend mean\_volume\_mentions mean\_pos\_prop  
## <fct> <fct> <dbl> <dbl> <dbl>  
## 1 Camaro 0 71.1 87596. 0.244  
## 2 Camaro 1 518. 93613. 0.247  
## # ℹ 1 more variable: mean\_neg\_prop <dbl>

# Bivariate statistics:   
superbowl %>% filter(new\_brand=="Camaro") %>% group\_by(new\_brand, superbowl) %>% summarise(cor\_adspend\_and\_volume=cor(adspend,volume), cor\_adspend\_and\_pos\_prop=cor(adspend,pos\_prop), cor\_adspend\_and\_neg\_prop=cor(adspend,neg\_prop))

## `summarise()` has grouped output by 'new\_brand'. You can override using the  
## `.groups` argument.

## # A tibble: 2 × 5  
## # Groups: new\_brand [1]  
## new\_brand superbowl cor\_adspend\_and\_volume cor\_adspend\_and\_pos\_prop  
## <fct> <fct> <dbl> <dbl>  
## 1 Camaro 0 -0.104 0.429  
## 2 Camaro 1 0.115 0.121  
## # ℹ 1 more variable: cor\_adspend\_and\_neg\_prop <dbl>

##### The univariate statistics return the mean ad spend for Camaro, which can then be compared to the other columns returned in the query, like mean volume of mentions and the mean positive and negative proportion of mentions. Based on the results of this first query, we can see that the mean ad spend for Camaro increased by $446,900, from an average of $71,100 to $518,000 - before and after the Super Bowl, respectively. With this average increase in ad spend, the average volume of mentions also increased by 6,017, from an average volume of 87,596 to 93,613. The average proportion of positive mentions also stayed the same, only increasing by 0.003, from 0.244 to 0.247, meaning almost a quarter of the mentions before and after the Super Bowl were positive. On average, only about five percent of mentions were negative before and after the Super Bowl, with the average proportion of negative mentions actually decreasing by 0.0023.

##### The bivariate statistics return the correlation between adspend and volume, pos\_prop, and neg\_prop for Camaro. The correlation between ad spend and volume was negative, but not too substantial before the Super Bowl. After the Super Bowl, this correlation became positive, yet it was not very significant either. Before the Super Bowl, the correlation between ad spend and both pos\_prop and neg\_prop both show a moderate, positive correlation, which implies that as ad spend increases, both the proportion of positive and negative mentions tends to increase as well. After the Super Bowl, there are very weak, positive correlations between ad spend and volume, pos\_prop, and neg\_prop. Since the correlation between ad spend and volume, pos\_prop, and neg\_prop after the Super Bowl were very close to 0, I would not rule them as having a significant correlation.

##### Considering these insights, while ad spending may contribute to creating social media buzz, the effectiveness varies across brands and might not solely rely on ad expenditure. It’s essential to evaluate other factors influencing online engagement and brand perception. Investing millions in ads solely for social media buzz might not guarantee desired outcomes, as seen in this analysis of brand-specific metrics and correlations.