Super Bowl Advertisement Data Analysis

By: Reagan Newswanger

CSC558: Dr. Parson

May 2023

4.2. a Source of Data

- o   Source of Data: Kaggle
- o   URL: https://www.kaggle.com/datasets/thedevastator/uncover-america-s-secrets-through-super-bowl-ads
- o   The data was collected from data.world's Admin.

4.2. b Goal

I am starting a new analysis. I am taking a business approach in analyzing the data. By analyzing this data, I hope to determine which companies have had the best success when it comes to Super Bowl commercials and what attributes play a role in the success. Analyzing trends could help companies in the future when planning on what type of advertisement they would like to use to increase popularity.

4.2.c Data File Steps

I downloaded the Excel file from Kaggle. I analyzed it and realized some YouTube URLs were missing. Any data entry that was missing data was deleted. I also used all the YouTube URLs and viewed how many views each video obtained and created a new column labeled youtube_views. Once the data was clean and I added all the YouTube views I saved the Excel file as a CSV file and uploaded it to Weka. Once the CSV file was uploaded to Weka, I removed the superbowl_ads_dot_com_url and youtube_url attributes because there is no use for them to analyze the data. Once those attributes were removed, I saved the file as a .ARFF file.

4.2.d Commercial or Research Setting

The results could be studied by corporations to better understand what attributes tend to lead to higher views and popularity. A corporation could analyze its own advertisements throughout the years and analyze its competitors and use that data to improve its popularity of Super Bowl advertisements in the future. Corporations spend millions of dollars on Super Bowl Advertisements. It is beneficial to add analytics to the creative process of marketing to achieve the best possible results in an advertisement's success.

4.2.e Technique Anticipated to be Used

My original thought was that I was going to use Weka as my modeling tool, but I might end up using Excel to get the results I am looking for. I possibly may use the combination of Excel and Weka. My data consists of qualitative and quantitative data so my techniques will vary depending on what attribute I am analyzing. It seems like Excel will give better results when referring to 4.2.d., but I will not be sure until I start the analyzing process.

4.2.f Document any other aspect of the project that you feel is important to communicate

Currently I do not have any specifics about the project that I feel is necessary to communicate about.

5.2.a Additional Data

I did not collect any additional data after assignment 4 was completed. As mentioned in assignment 4 I added the YouTube views to each data entry. The YouTube URL's can be found on the original data set.

5.2b Was Goal Achieved

My intended goal for this data was to be able to determine which companies have had the most success when it comes to Super Bowl commercials and what attributes play a role in the success. By being able to pull certain data by using PivotTables and then analyzing that information with regression analysis it resulted in being able to see which companies are the were the most popular and which attributes lead to their successful commercials. I created a PivotTable showing each brand and the number of views they each got. See below:

brand	Sum of youtube_views	Percentage of Views
Doritos	209098300	65.16%
Coca-Cola	44491510	13.87%
Budweiser	38474590	11.99%
Bud Light	14121299	4.40%
NFL	5838000	1.82%
Pepsi	3573194	1.11%
Hynudai	2017758	0.63%
E-Trade	1965749	0.61%
Toyota	880100	0.27%
Kia	422267	0.13%
Grand Total	**320882767**	**100%**

The correlation coefficient for this data was 0.6868 and sig f was .0282. This shows that there is a medium to strong correlation between brand and number of views. Doritos dominated and accounted for 65.16% of the total views. Doritos, Coca-Cola, and Budweiser collectively account for over 90% of the views so I decided to focus on those three companies for the majority of the anakysis. See the top 3 brand's information below.

Top 3 Brands with the Most Views:

Brand	Top Category	Second Category	Third Category	Correlation Coefficient	Sig F
1.Doritos	Funny	Animal	Danger	0.6935	0.1265
2.Coca-Cola	Danger	Animal	Funny	0.9710	0.0012
3.Budweiser	Patriotic	Danger	Funny	0.9108	0.0115

Top 3 Categories for each Brand:

Brand (In order by views)	Top Category	Second Category	Third Category
1.Doritos	Funny	Animal	Danger
2.Coca-Cola	Danger	Animal	Funny
3.Budweiser	Patriotic	Danger	Funny
4.Bud Light	Funny	Danger	Animals
5.NFL	Celebrity	Patriotic	Funny
6.Pepsi	Celebrity	Funny	Use_Sex
7.Hyundai	Celebrity	Animals	Funny
8.E-Trade	Funny	Use_Sex	Patriotic
9.Toyota	Funny	Danger	Animals
10.Kia	Funny	Use_Sex	Celebrity

As seen above in the "Top 3 Brands with the Most Views" chart all had funny and danger as a top category and two out of the three had animals in the top categories. This indicates that funny, danger, and animals lead to millions of views.

I added the "Top 3 Categories for each Brand" chart for reference to be able to see which categories were in the top 3 for each brand. I did not run regression for every brand. The only category that was within the top three for each brand was funny. So, while brand 4-10 had lower views overall, the views the brands do have success with were because funny was involved in the advertisements.

Overall, it is a good idea for any brand to always include funny aspects in the Super Bowl advertisements because it has a high success rate with views. The top three brands dominated and by far had the greatest number of views and all had the same categories except one category. Because of this other companies should focus on including the funny, danger, and animal categories in their advertisements if they don't have them included already.

5.2.c Steps

I used Excel to analyze the Super Bowl Advertisements. I was unable to manipulate the data in Weka resulting in me utilizing excel for most of the analyzation. Excel allowed me to manipulate the data as I needed. I created a lot of PivotTables which allowed me to

analyze only portions of the data instead of each attribute. I ran a lot of linear regressions to view the correlation coefficient and sig F. Below are the specific steps I took:

The first step I took was analyzing the brands by how many total YouTube views they each got because views indicate the popularity of the commercial advertisements. There could be some flaws in basing popularity off the views, but overall, it is a good indicator to assist analyzing this data. I created a PivotTable to be able to pull how many views each brand had. Once the PivotTable was created, I ran regression analysis from the data the PivotTable pulled. Below shows PivotTable and the regression results.

Regression Statistics	
Multiple R	0.686817314
R Square	0.471718022
Adjusted R Square	0.405682775
Standard Error	49523525.64
Observations	10

brand	Sum of youtube_views	Percentage of Views
Doritos	209098300	65.16%
Coca-Cola	44491510	13.87%
Budweiser	38474590	11.99%
Bud Light	14121299	4.40%
NFL	5838000	1.82%
Pepsi	3573194	1.11%
Hynudai	2017758	0.63%
E-Trade	1965749	0.61%
Toyota	880100	0.27%
Kia	422267	0.13%
Grand Total	**320882767**	**100%**

ANOVA

	df	SS	MS	F	Significance F
Regression	1	1.75198E+16	1.75198E+16	7.143427827	0.028242939
Residual	8	1.96206E+16	2.45258E+15		
Total	9	3.71405E+16			

The two focal points of the regression output are Multiple R (correlation coefficient) and Significance F which states if the null hypothesis is rejected or accepted. This Sig F accepted the Alternative Hypothesis which means the linear regression model is significant.

Next, I created a PivotTable to be able to see how many views each category had. I ran a regression with the PivotTable results.

category	Number of Views
use_sex	308998252
patriotic	280374496
animals	263105881
funny	252960992
danger	234875166
celebrity	32331008
Grand Total	**228774299**

Regression Statistics	
Multiple R	0.822017013
R Square	0.675711969
Adjusted R Square	0.594639962
Standard Error	1.191117179
Observations	6

ANOVA

	df	SS	MS	F	Significance F
Regression	1	11.82495946	11.824959	8.334713654	0.044697848
Residual	4	5.675040537	1.4187601		
Total	5	17.5			

After I analyzed the categories as a whole, I then analyzed the top 3 brands to see which category they had the most views in. Once I saw the results of the brands ranked in order it was clear that Doritos substantially had the highest number of views and accounted for 65.16% of the views. I decided to only analyze the top 3 brands because they account for over 90% of the views. Below is the breakdown and analysis of Doritos, Coca-Cola, and Budweiser.

Doritos:

Regression Statistics	
Multiple R	0.693500007
R Square	0.48094226
Adjusted R Square	0.351177825
Standard Error	1.506943136
Observations	6

Doritos Topic	Sum of # of Views
funny	209,011,300
animals	15,284,000
danger	5,253,000
use_sex appeal	4,960,600
celebrity	2,096,300
patriotic	1,300,000
Grand Total	**237905200**

ANOVA

	df	SS	MS	F	Significance F
Regression	1	8.41649	8.416489545	3.706271749	0.12651672
Residual	4	9.08351	2.270877614		
Total	5	17.5			

Coca-Cola:

Regression Statistics	
Multiple R	0.971011374
R Square	0.942863089
Adjusted R Square	0.928578861
Standard Error	0.499973986
Observations	6

Coca-Cola Topic	Sum of # of Views
danger	41,382,000
animals	27,812,000
funny	17,939,700
celebrity	16,875,000
patriotic	373,000
use_sex appeal	810
Grand Total	**104382510**

ANOVA

	df	SS	MS	F	Significance F
Regression	1	16.5001	16.50010405	66.00728438	0.00124833
Residual	4	0.999896	0.249973987		
Total	5	17.5			

Budweiser:

Regression Statistics	
Multiple R	0.910893572
R Square	0.8297271
Adjusted R Square	0.787158875
Standard Error	0.863101347
Observations	6

Budweiser Topic	Sum of # of Views
patriotic	31,495,021
danger	28,857,800
funny	6,426,366
animals	5,229,721
use_sex appeal	1,388,783
celebrity	189,286
Grand Total	**73586977**

ANOVA

	df	SS	MS	F	Significance F
Regression	1	14.52022	14.52022426	19.4917007	0.011556183
Residual	4	2.979776	0.744943936		
Total	5	17.5			

Next, I ran Rank and Percentile within the Toolkit to see which years had the most views. Below are the results.

Point	year	Rank	Percent	Point	Sum of youtube_views	Rank	Percent
21	2020	1	100.00%	13	212625601	1	100.00%
20	2019	2	95.00%	18	28320000	2	95.00%
19	2018	3	90.00%	17	24258511	3	90.00%
18	2017	4	85.00%	15	8964400	4	85.00%
17	2016	5	80.00%	20	8447786	5	80.00%
16	2015	6	75.00%	1	4943600	6	75.00%
15	2014	7	70.00%	16	4427000	7	70.00%
14	2013	8	65.00%	8	4194454	8	65.00%
13	2012	9	60.00%	10	4002200	9	60.00%
12	2011	10	55.00%	9	3802400	10	55.00%
11	2010	11	50.00%	14	3138569	11	50.00%
10	2009	12	45.00%	21	2881800	12	45.00%
9	2008	13	40.00%	5	2561618	13	40.00%
8	2007	14	35.00%	19	1981510	14	35.00%
7	2006	15	30.00%	11	1979966	15	30.00%
6	2005	16	25.00%	7	1266000	16	25.00%
5	2004	17	20.00%	3	1058100	17	20.00%
4	2003	18	15.00%	12	860058	18	15.00%
3	2002	19	10.00%	4	679600	19	10.00%
2	2001	20	5.00%	2	431600	20	5.00%
1	2000	21	0.00%	6	57994	21	0.00%

Below is the regression for the above data.

Regression Statistics	
Multiple R	0.148902
R Square	0.022172
Adjusted R Square	-0.02929
Standard Error	6.295059
Observations	21

ANOVA

	df	SS	MS	F	Significance F
Regression	1	17.07233335	17.07233	0.430817392	0.519459069
Residual	19	752.9276667	39.62777		
Total	20	770			

The regression shows a very week correlation coefficient and has a sig f of .5129 which means the null is accepted and that the linear model is not significant.

Problems:

As mentioned earlier I was unable to use Weka as I thought I would be able to, so I had to switch my tool to Excel. Within Excel there are limited options of what data analysis the Analysis ToolPak can run. This limited the test I could run. Overall, I mostly used regression because the other tests were not useful to analyzing the trends of the Super Bowl Advertisement data. If I had more time or were to revisit this project I would research to see if there were additional tests or ways, I could efficiently analyze this data.

5.2.d Technique Not Used in Previous Projects

I did not use SMO, SMOreg, MultiLayerPerceptron, or clustering since I used Excel and not Weka. I know PivotTables are not considered machine learning, but they helped me to be able to manipulate the data the way I needed to be able to analyze the data. The techniques I used that were different than previous assignments are Rank and Percentile, t-Test, and Anova: Single. In my opinion the t-Test and Anova: did not provide me with data that was beneficial to achieving the goal of my analysis, so I stopped running the two tests. Below are examples of both tests that I ran.

t-Test: Paired Two Sample for Means:

	TRUE	FALSE
Mean	25296099.2	6792178
Variance	4.20625E+15	1.45E+14
Observations	10	10
Pearson Correlation	-0.143809976	
Hypothesized Mean Difference	0	
df	9	
t Stat	0.864953301	
P(T<=t) one-tail	0.204763342	
t Critical one-tail	1.833112933	
P(T<=t) two-tail	0.409526684	
t Critical two-tail	2.262157163	

Anova: Single

SUMMARY

Groups	Count	Sum	Average	Variance
brand	10	55	5.5	9.166667
Count of brand	10	213	21.3	186.0111

ANOVA

Source of Variation	SS	df	MS	F	P-value	F crit
Between Groups	1248.2	1	1248.2	12.79039	0.002158	4.413873
Within Groups	1756.6	18	97.58889			
Total	3004.8	19				

5.2.e Revise: Commercial or Research Setting

My original thought on this remains the same. There may be some flaws in the analysis, but this could be a starting point for businesses to analyze and they can build off the data and add more detailed attributes as needed.

The results could be studied by corporations to better understand what attributes tend to lead to higher views and popularity. A corporation could analyze its own advertisements throughout the years and analyze its competitors and use that data to improve its popularity of Super Bowl advertisements in the future. Corporations spend millions of dollars on Super Bowl Advertisements. It is beneficial to add analytics to the creative process of marketing to achieve the best possible results in an advertisement's success.

5.2.f Important to Communicate

Due to the time constraint of this project, I believe there may be aspects of the data that I have not analyzed that could be beneficial to analyze or take into consideration. I did the best I could with the amount of time I had. For example, each brand did not have the same number of advertisements and I did not take that into account when running numbers. I do wish I had more time to analyze this information. But I do want to note that Doritos did not have the greatest number of advertisements and Kia, which has the lowest number of views, does not have the least number of advertisements. The breakdown is provided for reference.

brand	Count of brand
Bud Light	50
Budweiser	37
Doritos	24
Pepsi	24
Hynudai	20
Coca-Cola	20
E-Trade	13
Kia	11
Toyota	8
NFL	6
Grand Total	213

I believe with the data that was included in the original excel file and the time frame I had; I did achieve the goal. But I recognize there could be factors that I didn't take into consideration that could affect the outcomes if I had more time to investigate different aspects of the project.