Stat 172: Project #1

Name: Samuel Tucker

Description: Modeling Binary Data to Create a Successful Video for MrBeast

Date: 9.30.22

1. Introduction

The data used in this analysis is called Mr Beast YouTube Video Statistics. The author Rob Mulla collected the data straight from MrBeast's YouTube page and uploaded the data to the website kaggle.com. As of the above date, this data was last updated December 20, 2021. The data was first cleaned in R. Steps were taken to subset the needed columns, delete rows with blank entries, and remove duplicate rows. The R-cleaned data was then exported and moved into SAS in order to finish cleaning. In SAS, a few outliers were removed, and transformations were performed onto columns to better interpret and manipulate those columns.

I chose this data because I've always been fascinated with the way content creators such as MrBeast grow their following *and maintain it*. The data lays out videos from as far back as 2013 when MrBeast, at the time, only had 30 subscribers. Today, MrBeast's YouTube channel has 105 million subscribers. My goal in this analysis is to be MrBeast's data analyst. MrBeast has a new idea for a video, so he comes to me and asks: based on previous videos that were successful, what are important variables to consider when attempting to make a new video?

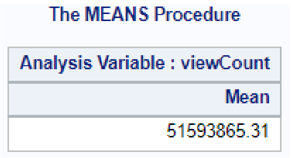
The following are included in this whitepaper:

- Defining and exploring all the variables used in the analysis
- Specifying and justifying the model that is used
- Giving helpful interpretations to the model's results
- Supplying the R and SAS code used for the analysis

1.1 Response Variable

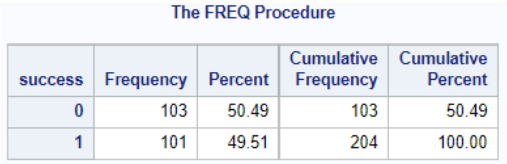
The response variable created from this dataset is called *success*. *Success* is binary and is derived from the column describing the total number of views a video has. If a video is viewed more than an average MrBeast video, *success* is 1. If a video is viewed less than an average MrBeast video, *success* is 0. See Figure 1 below.

Figure 1:



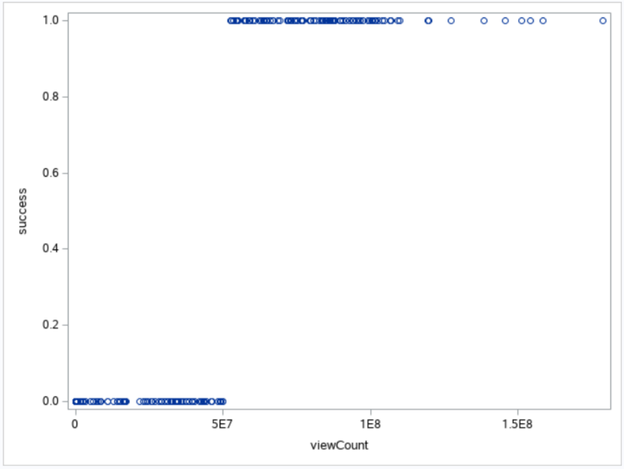
The MEANS Procedure

Analysis Variable : viewCount
Mean
51593865.31



The FREQ Procedure

success	Frequency	Percent	Cumulative Frequency	Cumulative Percent
0	103	50.49	103	50.49
1	101	49.51	204	100.00

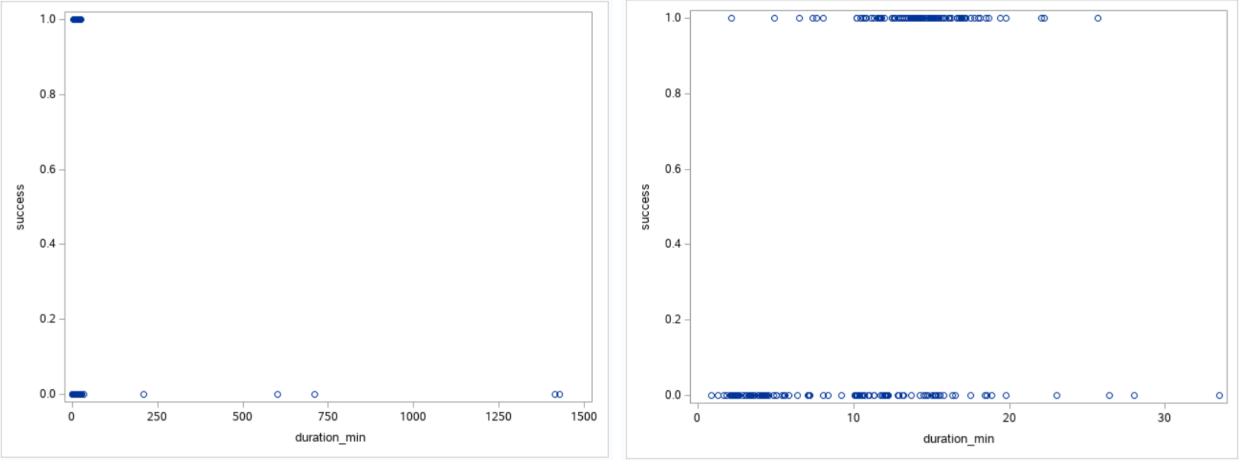


I expected the *success* variable to have a fairly even 50/50 split. The frequency table for the *success* variable (Fig 1. bottom left) proved this to be true. As expected in the scatterplot (Fig 1. right side), the *viewCount* values on the x-axis greater than the mean returned 1 for *success*. Similarly, the *viewCount* values on the x-axis less than the mean returned 0 for *success*. At this point, I'm confident that the response variable was created correctly, and I can now move on to the explanatory variables.

1.2 Explanatory Variables

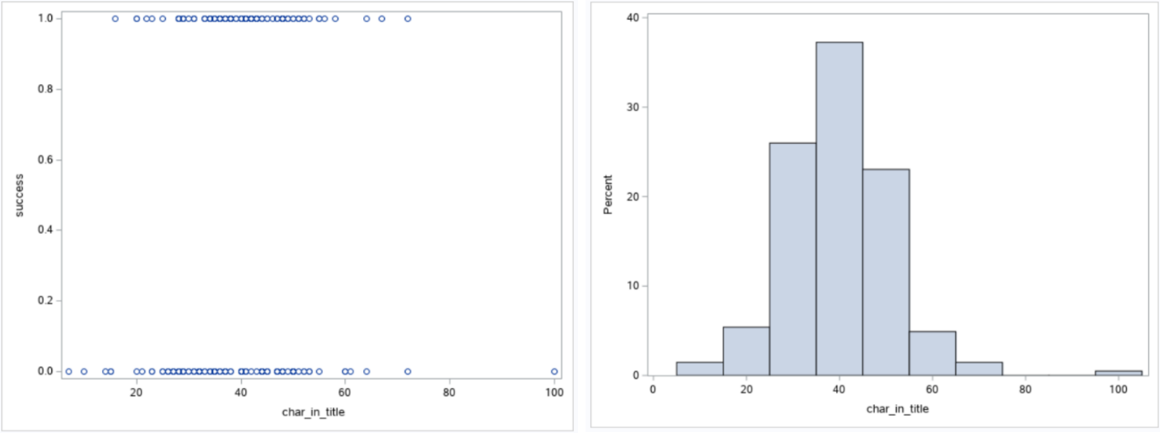
Four explanatory variables make up the final model. Two numeric. Two categorical. The first numerical variable is named *duration_min*. *Duration_min* is a transformation of the column *duration_seconds* where the values in *duration_seconds* are converted from seconds to minutes. The new *duration_min* variable had flaws, though. In Figure 2 below, outliers are abundant in the left scatterplot. All the outliers produced a 0 for *success*, and heavily skewed the *duration_min* variable. Overall, they were unhelpful. So, I removed them. In Figure 2's right scatterplot below, we find the outliers removed, resulting in a much cleaner plot.

Figure 2:



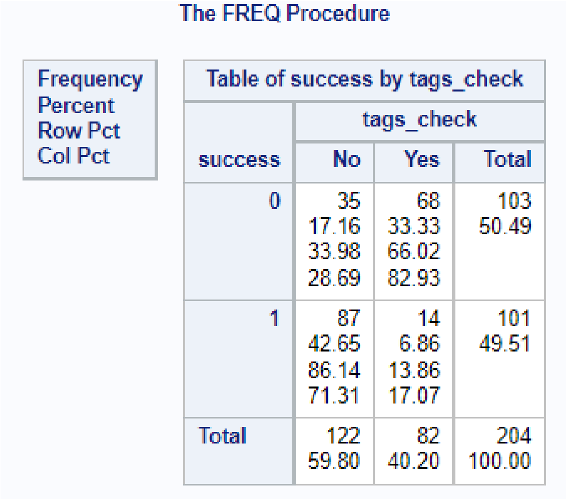
The second numerical variable is named *char_in_title*. This variable is created by summing the number of characters for each title. In Figure 3 below, we can see a scatterplot and histogram of the values in *char_in_title* in reference to the *success* variable.

Figure 3:



The first categorical variable is called *tags_check*. The purpose of this variable is to see if tags were used when the video was posted. If tags were used, the variable would read “Yes”. If tags were not used, the variable would read “No”. In Figure 4 below, you can see the disbursement of *tags_check* between the binary *success* variable. An interesting note is there is more *success* when tags are *not* used. Supposedly, tags are supposed to increase engagement leading to more views, but it seems here that that is not the case. More on this later.

Figure 4:



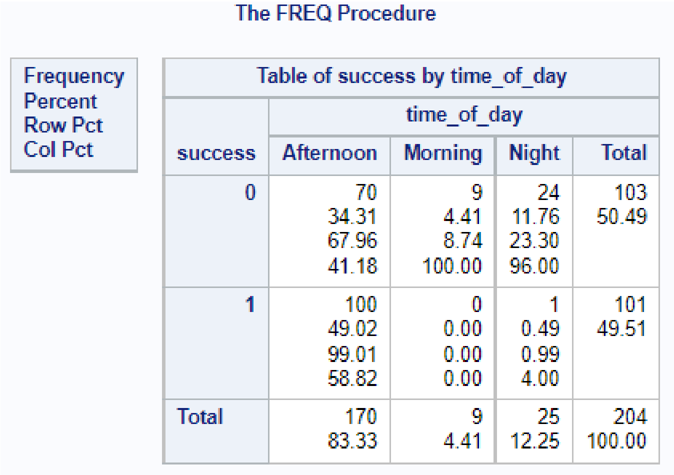
The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of success by tags_check			
	tags_check		
success	No	Yes	Total
0	35 17.16 33.98 28.69	68 33.33 66.02 82.93	103 50.49
1	87 42.65 86.14 71.31	14 6.86 13.86 17.07	101 49.51
Total	122 59.80	82 40.20	204 100.00

The second categorical variable and the last explanatory variable of our model is *time_of_day*. Using the *publishTime* column, *time_of_day* reads the hour of the day and categorizes it accordingly. 5am-noon (CDT) indicates “Morning” for *time_of_day*. Noon-6pm (CDT) indicates “Afternoon” for *time_of_day*. Finally, 6pm-5am (CDT) indicates “Night” for *time_of_day*. In Figure 5 below, we can see the disbursement of *time_of_day* in reference to our *success* variable. One immediate observation is 100/101 of the times *success* is 1 *time_of_day* indicates “Afternoon”. Another observation to note is that when *time_of_day* is “Morning”, *success* is never 1. This will indicate complete separation in our model, which we’ll discuss later.

Figure 5:



The FREQ Procedure

Frequency
Percent
Row Pct
Col Pct

Table of success by time_of_day				
	time_of_day			
success	Afternoon	Morning	Night	Total
0	70 34.31 67.96 41.18	9 4.41 8.74 100.00	24 11.76 23.30 96.00	103 50.49
1	100 49.02 99.01 58.82	0 0.00 0.00 0.00	1 0.49 0.99 4.00	101 49.51
Total	170 83.33	9 4.41	25 12.25	204 100.00

2. Methods

2.1 Specification

$$success_i = \begin{cases} 1, if\ viewCount\ is\ greater\ than\ the\ average\ of\ viewCount \\ 0, otherwise \end{cases}$$

Also included,

- Let $duration\_min_i$ represent the duration in minutes of a video *i*.
- Let $char\_in\_title_i$ represent the total characters in the title of a video *i*.
- "Yes" is the baseline category of $tags\_check_i$ and represents the presence of tags for a video *i*. *S*o, the other level will have a variable in the linear predictor.
  - Let $NO\_tags\_check_i$ = 1 if there is not a presence of tags for a video *i*.
- "Night" is the baseline category of $time\_of\_day_i$ and represents the time of day a video *i* is posted. So, the 2 other levels will have a variable in the linear predictor.
  - Let $Afternoon\_time\_of\_day_i$ = 1 if the time of posting a video *i* is in the afternoon.
  - Let $Morning\_time\_of\_day_i$ = 1 if the time of posting a video *i* is in the morning.

Random Component:

$$success_i \sim Bernoulli(\pi_i)$$

Systematic Component:

- Link Function:

$$g(\pi_i) = \eta_i = \log\left(\frac{\pi_i}{1-\pi_i}\right)$$

- Linear Predictor

$$\eta_i = \beta_0 + \beta_1 duration\_min_i + \beta_2 char\_in\_title_i + \beta_3 NO\_tags\_check_i + \beta_4 Afternoon\_time\_of\_day_i + \beta_5 Morning\_time\_of\_day_i$$

2.2 Support for the Model

The generalized linear model (see above) is appropriate for our analysis because $success_i$ is binary and fluctuates between 1 and 0 exclusively. Defined more concisely as Ω = $\{0,1\}$. Because $success_i$ is binary, it would make the most sense for this random variable to be Bernoulli. In addition, it is pertinent that a link function connects π_i to the linear predictor, because the identity link and the OLR model cannot. This is for a few reasons. We would assume normality for an OLR model, but this assumption is not met since a normal distribution is continuous and $success_i$ is binary. Additionally, since $success_i \sim Bernoulli(\pi_i)$, there cannot be a constant variance because the variance ($Var(Y_i) = \pi_i(1-\pi_i)$) will change based on the explanatory variables and coefficients. Finally, the identity link permits $\pi_i > 1$ and $\pi_i < 0$, which are impossible probabilities.

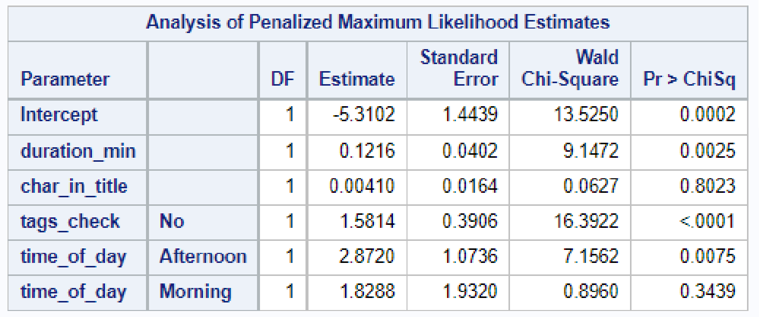
3. Results

3.1 Complete Separation

Complete separation can arise often when performing statistical modeling. This means that for an explanatory variable, it did not give a specific outcome given its observations. Simply speaking, the x variable perfectly predicts the y variable. The explanatory variable *time_of_day* contains complete separation. As we saw earlier in Figure 5, the cell in row 2, column 2 only has 0's. This explanatory variable has 0 observations in which a video posted in the morning produced a 1 for *success*. Despite the lack of observations in this cell, we will keep this variable in our model because the rest of the data from this variable is quite helpful. So, how does this affect us? It means our maximum likelihood estimate (MLE) does not exist. The MLE is how we interpret our x variables. The solution is "Firth's Penalized Likelihood." We add this to our code and use it in place of the MLE.

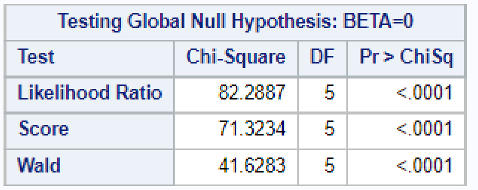
3.2 Interpretations and Confidence Intervals

Figure 6:



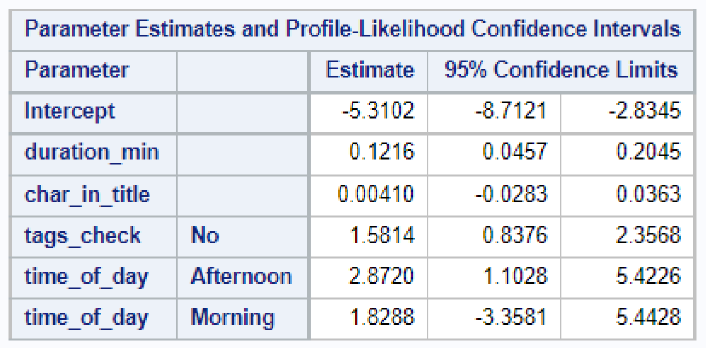
Analysis of Penalized Maximum Likelihood Estimates

Parameter		DF	Estimate	Standard Error	Wald Chi-Square	Pr > ChiSq
Intercept		1	-5.3102	1.4439	13.5250	0.0002
duration_min		1	0.1216	0.0402	9.1472	0.0025
char_in_title		1	0.00410	0.0164	0.0627	0.8023
tags_check	No	1	1.5814	0.3906	16.3922	<.0001
time_of_day	Afternoon	1	2.8720	1.0736	7.1562	0.0075
time_of_day	Morning	1	1.8288	1.9320	0.8960	0.3439



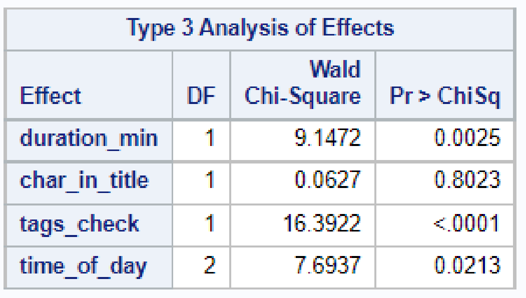
Testing Global Null Hypothesis: BETA=0

Test	Chi-Square	DF	Pr > ChiSq
Likelihood Ratio	82.2887	5	<.0001
Score	71.3234	5	<.0001
Wald	41.6283	5	<.0001



Parameter Estimates and Profile-Likelihood Confidence Intervals

Parameter		Estimate	95% Confidence Limits	
Intercept		-5.3102	-8.7121	-2.8345
duration_min		0.1216	0.0457	0.2045
char_in_title		0.00410	-0.0283	0.0363
tags_check	No	1.5814	0.8376	2.3568
time_of_day	Afternoon	2.8720	1.1028	5.4226
time_of_day	Morning	1.8288	-3.3581	5.4428



Type 3 Analysis of Effects

Effect	DF	Wald Chi-Square	Pr > ChiSq
duration_min	1	9.1472	0.0025
char_in_title	1	0.0627	0.8023
tags_check	1	16.3922	<.0001
time_of_day	2	7.6937	0.0213

Penalized Maximum Likelihood Estimates Interpretations:

β_1: Holding all other explanatory variables constant, the odds of a video being successful is predicted to increase by a factor of $e^{0.1216}$ = 1.1293 (or increase by ~12.93%) for every 1-minute increase in *duration_min*.

- For a video that with a 3-minute increase in *duration_min*, the odds of the video being successful is predicted to increase by a factor of $e^{3*0.1216}$ = 1.44 (or ~44%).

β_3: Holding all other explanatory variables constant, the odds of a video being successful is predicted to increase by a factor of $e^{1.5814}$ = 4.8618 (or increase by ~386.18%) for videos with no tags.

Parameter Estimates and Profile-Likelihood Confidence Intervals:

β_1: We are 95% confident that the odds of a video being successful change by a factor between $(e^{0.0457}, e^{0.2045})$ = (1.04676, 1.22691) for every 1-minute increase in *duration_min*.

- For a video that with a 3-minute increase in *duration_min*, the odds of the video being successful change by a factor between $(e^{3*0.0457}, e^{3*0.2045})$ = (1.14694, 1.84688).

β_3: We are 95% confident that the odds of a video being successful change by a factor between $(e^{0.8376}, e^{2.3568})$ = (2.31081, 10.55711) for videos with no tags.

The results of β_3 are unique. In the penalized MLE, we see that a video has increased success by ~385% if the video has *no* tags. Tags are, supposedly, supposed to help increase exposure for a posted video and would therefore increase views, leading to a greater chance of our definition of success. In addition, the 95% confidence interval for β_3 ranges from ~231% to ~1056%. Simply speaking, don't use tags! I think this is a huge finding, because it contradicts the popular opinion that you should always use tags when you post a video!

3.3 Hypothesis Testing

Testing significance of the entire model:

$$H_0: \beta_1 = \beta_2 = \beta_3 = \beta_4 = \beta_5 = 0$$
$$H_a: At\ least\ one\ \beta\ in\ \beta_1, \dots, \beta_5\ \neq 0$$

- Null distribution: $\chi^2(5)$ (Chi-square distribution with 5 degrees of freedom)
- Test statistic: *lr* = 82.2887
- p-value: < 0.0001
- Conclusion: Because the p-value is less than α = 0.05, we reject H_0. Further, we have enough evidence to say that at least one β is not equal to zero and our model as a whole is significant.

Individual hypothesis tests:

$$H_0: \beta_1 = 0$$
$$H_a: \beta_1 \neq 0$$

- Null distribution: $\chi^2(1)$ (Chi-square distribution with 1 degree of freedom)
- Test statistic: ω = 9.1472
- p-value: 0.0025
- Conclusion: Because the p-value is less than α = 0.05, we reject H_0. Further, we have enough evidence to say that $\beta_1 \neq 0$. This indicates that a video's duration in minutes has a significant relationship with the probability of the video receiving more than the average views.

$$H_0: \beta_2 = 0$$
$$H_a: \beta_2 \neq 0$$

- Null distribution: $\chi^2(1)$ (Chi-square distribution with 1 degree of freedom)
- Test statistic: ω = 0.0627

- p-value: 0.8023
- Conclusion: Because the p-value is greater than α = 0.05, we fail to reject H_0. Further, we do not have enough evidence to say that $\beta_2 \neq 0$. This indicates that a video's total characters in the title does not have a significant relationship with the probability of the video receiving more than the average views.

$$H_0: \beta_3 = 0$$
$$H_a: \beta_3 \neq 0$$

- Null distribution: $\chi^2(1)$ (Chi-square distribution with 1 degree of freedom)
- Test statistic: ω = 16.3922
- p-value: < 0.0001
- Conclusion: Because the p-value is less than α = 0.05, we reject H_0. Further, we have enough evidence to say that $\beta_3 \neq 0$. This indicates that a video that has no tags has a significant relationship with the probability of the video receiving more than the average views compared to a video that contains tags.

$$H_0: \beta_4 = \beta_5 = 0$$
$$H_a: At\ least\ one\ \beta_4, \beta_5 \neq 0$$

- Null distribution: $\chi^2(2)$ (Chi-square distribution with 2 degrees of freedom)
- Test statistic: ω = 7.6937
- p-value: 0.0213
- Conclusion: Because the p-value is less than α = 0.05, we reject H_0. Further, we have enough evidence to say that at least one $\beta_4, \beta_5 \neq 0$. This indicates that the time of day a video is posted has a significant relationship with the probability of the video receiving more than the average views.

$$H_0: \beta_4 = 0$$
$$H_a: \beta_4 \neq 0$$

- Null distribution: $\chi^2(1)$ (Chi-square distribution with 1 degree of freedom)
- Test statistic: ω = 7.1562
- p-value: 0.0075
- Conclusion: Because the p-value is less than α = 0.05, we reject H_0. Further, we have enough evidence to say that $\beta_4 \neq 0$. This indicates that the odds of a video receiving more than the average views is significantly different for a video posted in the afternoon versus a video posted at night.

$$H_0: \beta_5 = 0$$
$$H_a: \beta_5 \neq 0$$

- Null distribution: $\chi^2(1)$ (Chi-square distribution with 1 degree of freedom)
- Test statistic: ω = 0.8960
- p-value: 0.3439

- Conclusion: Because the p-value is greater than α = 0.05, we fail to reject H_0. Further, we do not have enough evidence to say that $\beta_5 \neq 0$. This indicates that we cannot say that the odds of a video receiving more than the average views is significantly different for a video posted in the morning versus a video posted at night.

3.4 Predictions

Let’s assume we have a video MrBeast wants to post with the following characteristics:

- The duration is 18 minutes.
- The title has 50 characters.
- The video is posted with no tags.
- The video is posted in the afternoon.

$$\hat{\eta} = -5.3102 + 0.1216 * 18 + 0.00410 * 50 + 1.5814 * 1 + 2.8720 * 1 + 1.8288 * 0$$

$$= -1.335$$

$$\hat{\pi} = 0.2632$$

Given these characteristics, we have a 26% probability of the video receiving more than the average views.

Now, let’s say we want to have a 70% probability of the video receiving more than the average views with all the same variables, but instead we solve for the duration in minutes.

$$\log\left(\frac{0.7}{1-0.7}\right) = -5.3102 + 0.1216 * duration\_min + 0.00410 * 50 + 1.5814 * 1 + 2.8720 * 1 + 1.8288 * 0$$

duration_min = 12.3281

Assuming the same variables in the first instance, if the video were to be edited down to 12.33 min, the video would have a 44% increased (70% - 26%) probability in receiving more than the average views. I think it is really helpful that changing one variable can affect our success so much.

4. Conclusion

In this whitepaper, we found that the duration, having no tags, and posting in the afternoon were significant variables when predicting if a video would have more than average views. Although our *time_of_day* variable contained complete separation, we were able to account for it and move forward. I think the biggest takeaway from this analysis was that the use of tags was insignificant. When a video had no tags, it did better. I think so often content creators think putting the right tags on a video will give them more exposure, but we see here that if anything, it hurts your exposure, almost limiting your exposure to those specific tags.

Further research could dive deeper into the definition of success. Another idea I had for creating this variable was creating a binary variable that looked at the previous video’s views and if the new video got more, the binary variable would be 1, and if it got less, then it would be 0. More simply, trying to answer the question: is my next video going to be better than the last one? Because as a content creator, the goal is to continue to get better and to grow. In addition, more explanatory variables could be used. Some of these might include diving deeper into the actual content of what a certain video was about. Was he giving away money? How much money? Was he eating exotic food? Was he doing a specific challenge that he’s done in the past but just in a different way? Was he reacting to another YouTube video? I’d be interested if any of these variables in future research were significant.

Appendix 1, R code:

```
# Name: Samuel Tucker
# Date: 9.26.22
# Description: cleaning Mr Beast YouTube data to eventually read into SAS

'Read in Mrbeast_youtube_stats'
beast = read.csv(choose.files(), header=T)

attach(beast)
head(beast)
colnames(beast)

# I want only these columns exclusively
beast_sub = subset(beast, select = c(title, publishTime, duration_seconds,
                                     viewCount, likeCount, commentCount,
                                     snippet.tags))

head(beast_sub)

#dropping na rows
dim(beast_sub)
beast_sub_good = na.omit(beast_sub)
dim(beast_sub_good)
# na.omit dropped 4 rows, good!

#removing duplicate data
library(dplyr)
beast_sub_good_clean = distinct(beast_sub_good)
dim(beast_sub_good_clean)
# 209 rows remain... good!

# time to export!
write.csv(beast_sub_good_clean, file = "Mrbeast_youtube_data_clean.csv", row.names = T)
```

Appendix 2, SAS code:

```
/*
Name: Samuel Tucker
Date: created: 9.26.22, finished: 9.30.22
Description: Code for project 1 Stat 172
*/

/*import the clean data that was exported from R*/
proc import out = beast
datafile = '/home/u57856520/Mrbeast_youtube_data_clean.csv'
dbms = csv replace;
guessingrows = max;
run;

/*creating duration_min before removing the outliers to show them*/
data beast;
set beast;
duration_min = round(duration_seconds/60, 0.01); /*60 seconds in a minute*/
run;

/*creating success variable for example*/
proc means data = beast mean;
   var viewCount;
run;
/*running this returns: 50859716.35 as the average*/

/*creating the success variable*/
data beast;
set beast;
success = viewCount;
if viewCount > 50859716.35 then success = 1;
        else success = 0;
run;

/*scatterplot with the outliers*/
proc sgplot data = beast;
scatter x = duration_min y = success;
run;
```

```
/*removing all data rows that have a duration_seconds > 9000. This data heavily skews this duration variable.*/
data beast;
set beast;
if duration_seconds < 9000;
run;

/*print imported data*/
proc print data = beast;
run;

/*checking that the removal of the outliers worked*/
proc freq data = beast;
tables duration_seconds;
run;

/*tasks:
        create numeric variable that counts characters in title
        create categorical variable that tells time of day (i.e. morning, afternoon, etc) from publishTime
        create binary variable that indicates if the video used a tag or not
*/

/*create binary variable that indicates if the video used tags or not*/
data beast;
set beast;
tags_check = snippet.tags;
if(length(snippet.tags) > 1) then tags_check = "Yes";
        else tags_check = "No";
run;

/*copying snippet.tags to tags to manipulate column better*/
data beast;
set beast;
tags = snippet.tags;
run;

/*checking that tags_check was created correctly*/
proc freq data = beast;
tables tags*tags_check;
```

```
run;
/*looks good!*/

/*creating categorical variable that tells what time of day the video was posted*/
data beast;
set beast;
time_of_day = publishTime;
time_whole_num = substr(publishTime, 11, 3);
run;

proc print data = beast;
run;

/*We live in the central timezone, so I will transfer the time_whole_num back 5 hrs*/

data beast;
set beast;
time_whole_num_right = input(time_whole_num, 8.);
time_whole_num_right_correct = time_whole_num_right - 5;
if (time_whole_num_right_correct < 0) then time_whole_num_right_correct = time_whole_num_right - 5 + 24;
run;

proc print data = beast;
run;

/*now to transfer time_whole_num_right_correct to a categorical variable*/
data beast;
set beast;
time_of_day = time_whole_num_right_correct;
if (time_whole_num_right_correct >= 5) & (time_whole_num_right_correct < 12) then time_of_day = "Morning";
	else if (time_whole_num_right_correct >= 12) & (time_whole_num_right_correct < 18) then time_of_day = "Afternoon";
	else if (time_whole_num_right_correct >= 18) & (time_whole_num_right_correct < 22) then time_of_day = "Evening";
	else if (time_whole_num_right_correct >= 22) then time_of_day = "Night";
	else if (time_whole_num_right_correct < 5) then time_of_day = "Night";
run;
```

```
/*checking that the time_of_day was distributed correctly*/
proc freq data = beast;
tables time_whole_num_right_correct*time_of_day;
run;

/*the categorical variable for time of day has been made for CDT using the data from publishTime!*/

/*Now what's left is creating a numerical variable based on the total number of characters in the title*/
data beast;
set beast;
char_in_title = length(title);
run;

/*checking variable type*/
proc contents data = beast;
run;

proc print data = beast;
run;

/*Next, I need to define a binary "success" variable based off # of views.
Since views directly impact watch time which impacts incoming revenue, I'm choosing # of views to
define success. I intend to take the average of the views column and define success as:
        1: a video having more views than the average
        0: a video having less views than the average
*/

/*first, let's compute the average of the viewCount column*/
proc means data = beast mean;
   var viewCount;
run;
/*running this returns: 51593865.31 as the average*/

/*creating success variable*/
data beast;
set beast;
success = viewCount;
if viewCount > 51593865.31 then success = 1;
        else success = 0;
```

```
run;

/*double check the logic works correctly*/
proc freq data = beast;
tables viewCount*success;
run;

/*table showing success variable details*/
proc freq data = beast;
tables success;
run;

/*scatterplot showing the viewCount*/
proc sgplot data = beast;
scatter x = viewCount y = success;
run;

/*done! our target variable and all of our explanatory variables are in place!
        success: binary variable
        duration_seconds: numerical variable
        tags_check: categorical variable
        time_of_day: categorical variable
        char_in_title: numerical variable
*/

/*checking that the above variable types are correct for each variable*/
proc contents data = beast;
run;

/*time to build our model!*/
/*after running once, was given */
proc logistic data = beast;
class time_of_day tags_check/ param = reference;
model success(event = '1') = duration_seconds tags_check time_of_day time_whole_num_right_correct
char_in_title/ clparm = both;
run;
/*complete separation might be detected... let's try a few things...*/

/*checking if this table returns an empty portion in the matrix*/
```

```
proc freq data = beast;
tables tags_check*success;
run;
/*false...*/
/*thinking it might be the time_of_day variable given the high standard error*/
proc freq data = beast;
tables time_of_day*success;
run;
/*yup!...*/
/*I will now add firth to the model*/
proc logistic data = beast;
class time_of_day tags_check/ param = reference;
model success(event = '1') = duration_seconds tags_check time_of_day time_whole_num_right_correct char_in_title/firth clparm = both;
run;
/*Was given this warning:
WARNING: Convergence was not attained in 25 iterations.

Did some light research. can we remove or re-define this time_of_day column?
re-defining it will be our solution
*/

/*A couple issues:
        - duration_seconds is alittle larger for our liking...
                - change this variable into minutes
        - time_of_day only has 1 observation for "Night"
                - I'm going to group together Night and Evening to just "Night" and see if that helps.
*/

/*First, changing duration_seconds to duration_min*/
data beast;
set beast;
duration_min = round(duration_seconds/60, 0.01); /*60 seconds in a minute*/
run;
/*done!*/

/*Now to group time_of_day better*/
data beast;
set beast;
```

```
if (time_whole_num_right_correct >= 5) & (time_whole_num_right_correct < 12) then time_of_day = "Morning";
	else if (time_whole_num_right_correct >= 12) & (time_whole_num_right_correct < 18) then time_of_day = "Afternoon";
	else if (time_whole_num_right_correct >= 18) then time_of_day = "Night";
	else if (time_whole_num_right_correct < 5) then time_of_day = "Night";
run;

/*double check this ran correctly*/
proc freq data = beast;
tables time_of_day*success;
run;

proc contents data = beast;
run;
/*Run our model again, these are our final variables
	success: binary variable
	duration_min: numerical variable
	char_in_title: numerical variable
	tags_check: categorical variable
	time_of_day: categorical variable
*/

/*Running our model again, this time using firth*/
proc logistic data = beast;
where duration_min < 150; /*removed 5 observations outliers.*/
class tags_check time_of_day/ param = reference;
model success(event = '1') = duration_min char_in_title tags_check time_of_day/firth clparm = both;
run;

/*various plots and tables used to show our variables*/
proc sgplot data = beast;
scatter x = duration_min y = success;
run;

proc freq data = beast;
tables duration_min*success;
run;
```

```
proc sgplot data = beast;
scatter x = char_in_title y = success;
run;

proc sgplot data = beast;
histogram char_in_title;
run;

proc freq data = beast;
tables success*tags_check;
run;

proc freq data = beast;
tables success*time_of_day;
run;

proc print data = beast;
run;
```