Youtube Analysis Final Report

Class: Statistics 112

Section: Discussion 1B

Group Number: 6

Members: Yechen Cao (806179375), Tishi Avvaru (506003820), Soomedha Vasudevan (805904117), Sun Moon Kim (205541619), Pavan Sah (806288100)

PowerPoint: [Youtube Analysis]

Table of Contents

Statement of the Problem (Research Question)

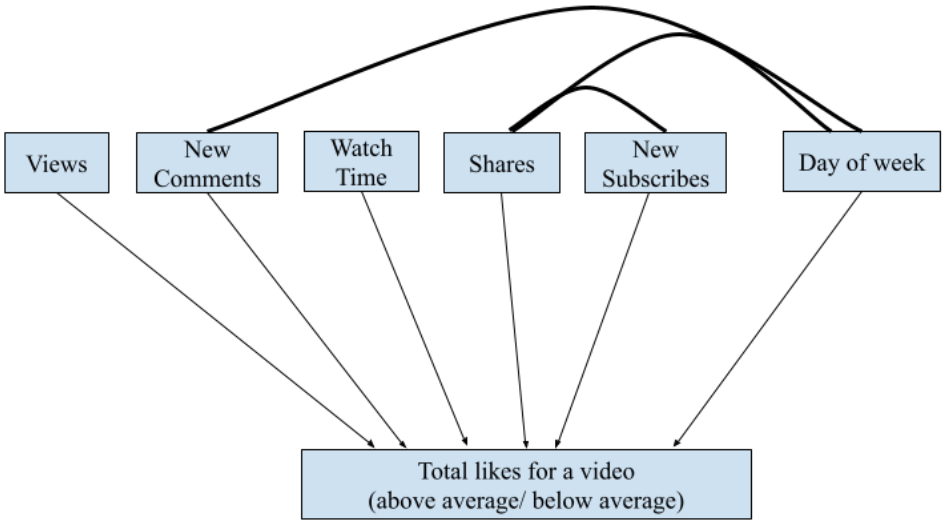


Figure 1: Path Diagram of the effects and interaction effects of explanatory variables on total likes for a video

Research Question:
Can the likelihood of a video getting above-average likes be predicted by metrics such as shares, new comments, watch time, new subscribers, views, and timing (day of week)?

Interaction Effects:
1. Does the effect of New Comments on the likelihood of a video getting above average likes vary with day of week?
2. Does the effect of Shares on the likelihood of a video getting above average likes vary with day of week?
3. Does the effect of Shares on the likelihood of a video getting above average likes vary with New Subscribers?

Abstract

This study investigates the key predictors that determine the likelihood of a YouTube video receiving above-average likes using a dataset of 364 videos. We performed logistic regression with forward selection to create our final model.

Our analysis indicated that shares, new subscribers, and watch time are the most significant in predicting the odds of a video receiving above-average likes. The final logistic regression model, after removing outliers, achieved an accuracy of 94.11% with an AIC of 108.17.

Conversely, variables such as new comments, day of the week, total views, and interactions between shares/day of the week and new comments/day of the week were not found to significantly increase the likelihood of receiving above-average likes.

Variables and Dataset

Outcome Variable:

- **Likes**: Binary variable indicating whether a video received above-average or below-average likes.

Predictor Variables:

1. **Number of Views** (Numerical) - Total views received by the video
2. **Number of new comments** (Numerical) - Number of new comments received on the video
3. **Number of Watch Time** (Numerical) - Total watch time in hours
4. **Number of shares** (Numerical) - Count of times the video was shared

5. **Number of new subscribers** (Numerical) - Number of new subscribers gained
6. **Day Type** (Categorical): Day of the week when the video was published

Dataset:

- Name: YouTube Channel Performance Analytics
- Size: 364

Exploratory Data Analysis (EDA)

Frequency Histogram

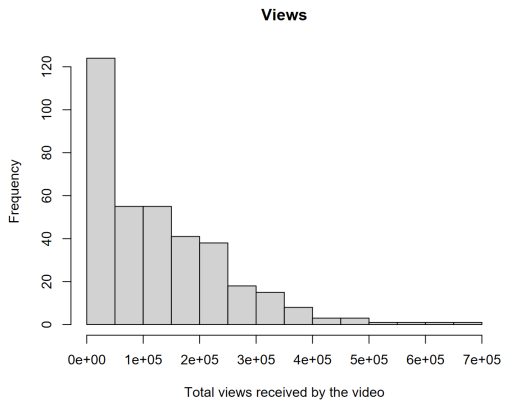


Figure 1: Frequency of view

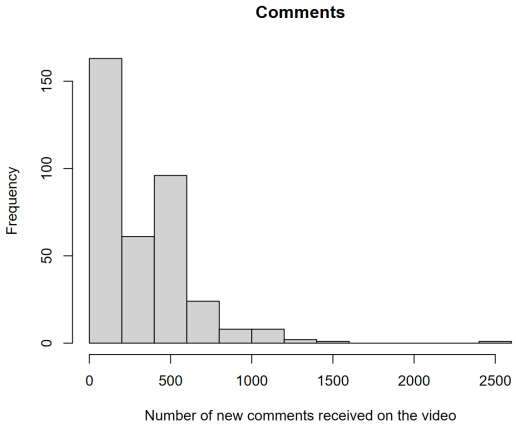


Figure 2: Frequency of new comments

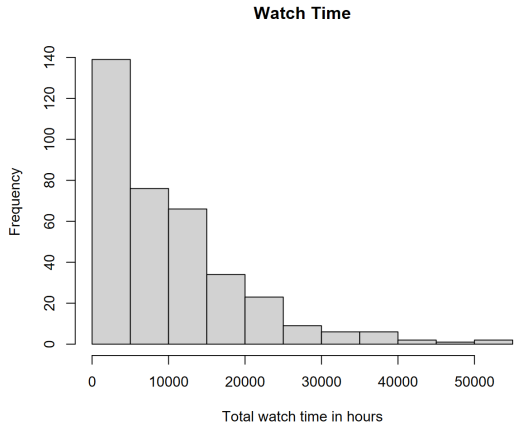


Figure 3: Frequency to watch time

The histograms illustrate the distribution of Views, New Comments, and Watch Time across the dataset. Figures 1 and 2 show that the majority of videos receive fewer than 100,000 views and fewer than 500 new comments, with both distributions being highly right-skewed. Similarly, Figure 3 reveals that most videos accumulate less than 10,000 hours of watch time, indicating a small subset of videos significantly outperform the rest.

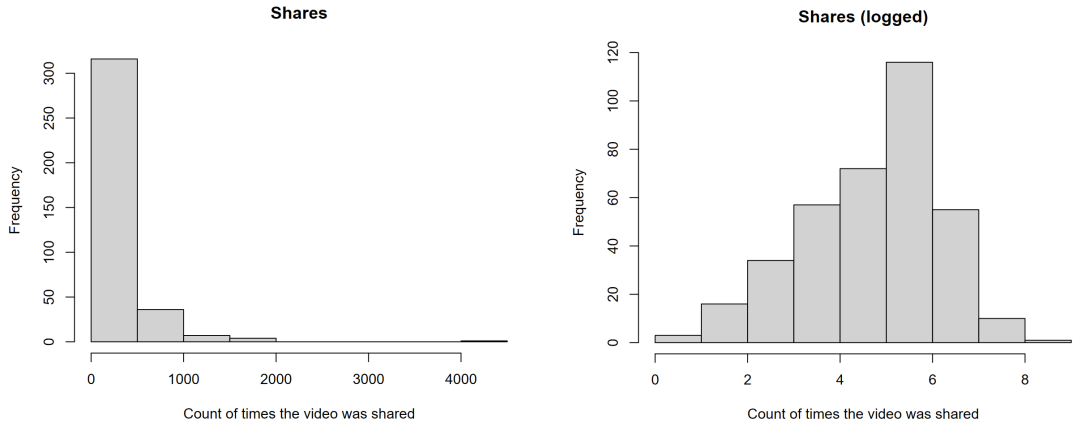


Figure 4: Frequency of Shares vs Shares(logged)

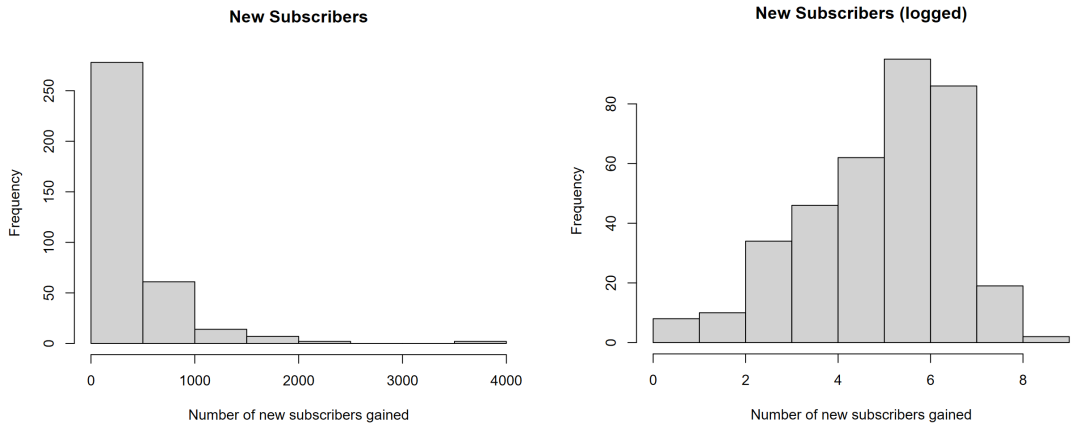


Figure 5: Frequency of New Subscribers vs New Subscribers(logged)

The histograms illustrate the distribution of Shares and New Subscribers across the dataset. Figures 4 and 5 reveal that the majority of videos are shared fewer than 1,000 times and attract fewer than 1,000 new subscribers. Both distributions are highly right-skewed, with a small subset of videos significantly outperforming the rest.

Log transformations were applied to both variables to address this skewness. The resulting histograms (right-hand side of Figures 4 and 5) display distributions closer to normal.

Correlation Heatmap

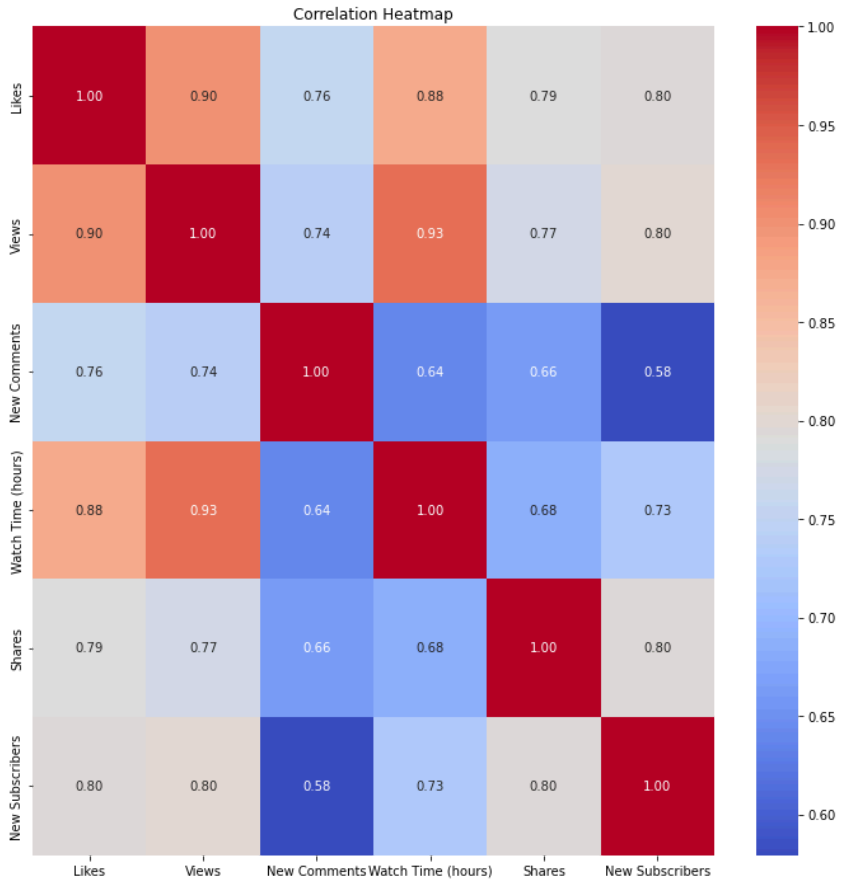


Figure 6: Correlation Heatmap

This correlation heatmap shows the relationships between the response variable, Likes, and the selected predictors. Focusing on the first row, which corresponds to the correlations with Likes, we can see that Views and Watch Time have the highest correlations, indicating a strong positive relationship with Likes. On the other hand, New Comments show the lowest correlation within these five variables. While it may appear relatively low on this heatmap, it is still higher than most variables in the full dataset, highlighting its significance. The heatmap effectively shows how each predictor contributes to the response variable, with all selected variables exhibiting positive correlations.

Scatter Plots

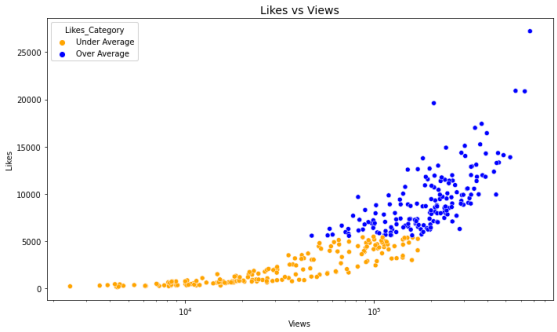


Figure 7: Likes vs. Views

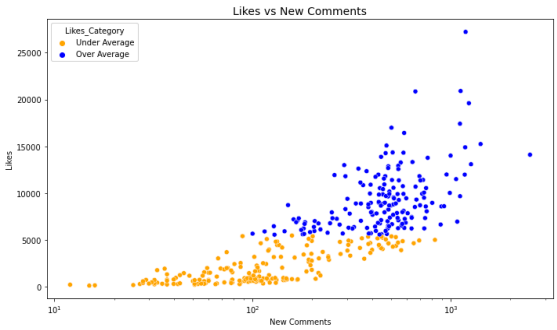


Figure 8: Likes vs. New Comments

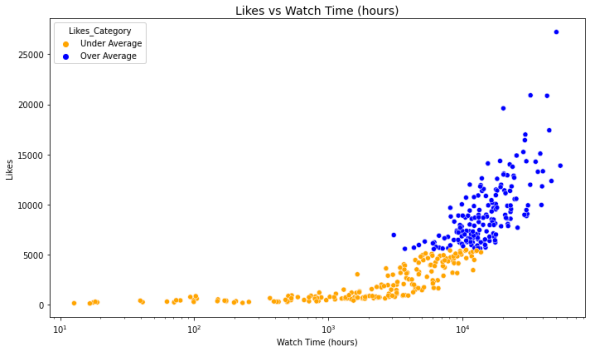


Figure 9: Likes vs. Watch Time (hour)

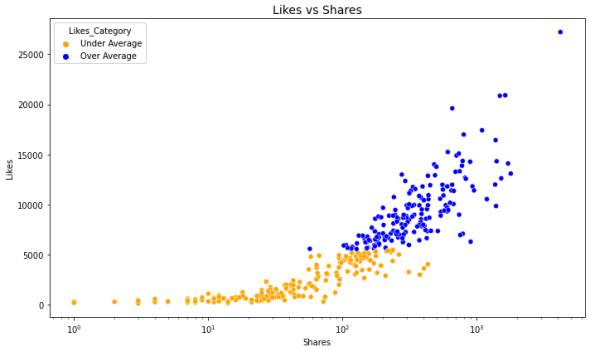


Figure 10: Likes vs. Shares

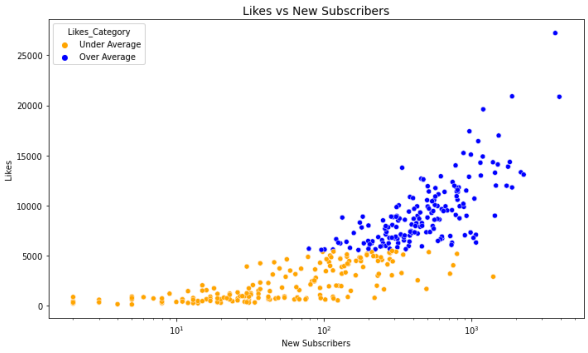


Figure 11: Likes vs. New Subscribers

The scatterplots show a positive correlation between likes and predictors such as Views, New Comments, Watch Time, Shares, and New Subscribers. Videos with above-average likes (blue points) generally have higher values for these predictors, highlighting their influence on engagement.

Bar Plots

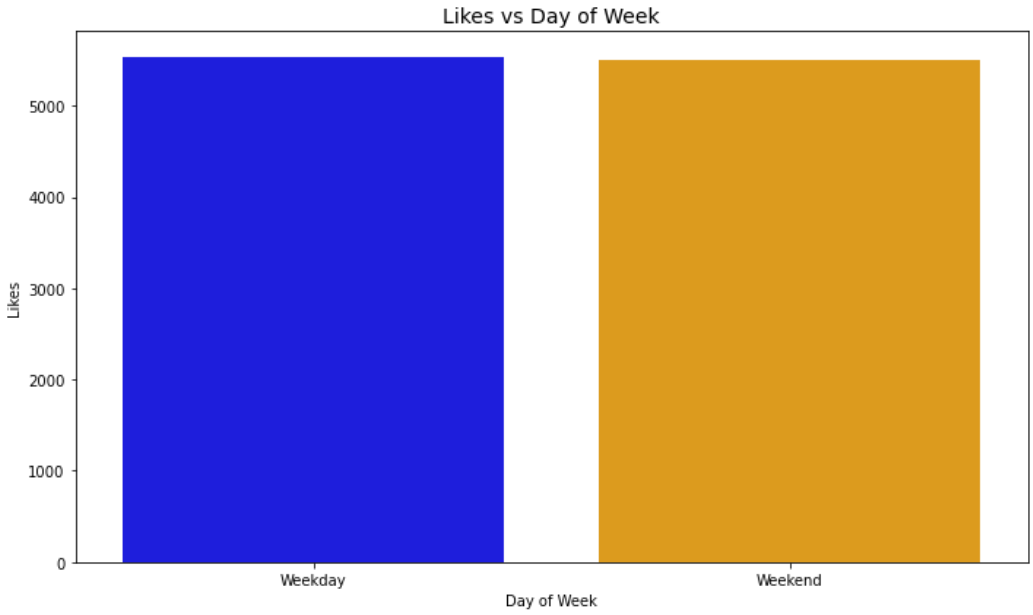


Figure 12: Likes vs. Day of Week

The bar plot shows the mean likes categorized by day type: weekdays and weekends. As shown in Figure 12, the average number of likes received on weekends is slightly higher than on weekdays.

Key Findings:

1. Typically, above average likes can be attributed to higher amounts of views, new comments, watch time, shares, and new subscribers.
2. Shares and new subscribers appear to have significant interaction. Type of day appears to be correlated with the variables.

Statistical Modeling and Assumption Checks

Linear Model:

- **Response Variable**: Likes.
- **Predictors**: Shares, New Comments, Views, Day Type, Subscribers

Assumption Checks:

1. **Residual vs Predictor**: No obvious patterns observed, indicating linearity.
2. **Residual vs Fitted**: Variance appeared homoscedastic.

Initial Model (Model 1)

Coefficients	Estimate	Std. Error	Z value	Pr(>|z|)
(Intercept)	-6.408e+00	9.039e-01	-7.089	1.35e-12 ***
New Comments	-1.009e-04	1.418e-03	-0.071	0.943274
Shares	1.472e-02	4.265e-03	3.450	0.000561 ***
New Subscribers	3.289e-03	1.077e-03	3.053	0.002266 **
Views	-1.245e-05	9.235e-06	-1.349	0.177489
Watch Time..hours	5.008e-04	1.053e-04	4.756	1.97e-06 ***
Day Type Weekend	-1.842e-01	1.360e+00	-0.135	0.892221
Shares: Day Type Weekend	-7.465e-03	6.378e-03	-1.170	0.241807
New Comments: Day TypeWeekend	8.891e-03	4.646e-03	1.914	0.055682
Shares: New Subscribers	-5.073e-06	1.692e-06	-2.998	0.002719 **
Null Deviance	503.72	Degrees of Freedom	363	
Residual Deviance	134.60	Degrees of Freedom	354	
AIC	154.6			
Number of Fisher Scoring Iterations	9			

Table 1: Summary for Initial Model

The initial model included predictors such as Shares, New Subscribers, Views, Watch Time, Day Type, and interaction terms. Significant predictors were Shares (p=0.000561), New Subscribers (p=0.002266), Watch Time (*p*=1.97×10^-6), and the interaction term Shares: New Subscribers

(p=0.002719). In contrast, predictors like New Comments, Views, and Day Type Weekend were not statistically significant. The model achieved a substantial reduction in deviance (134.60 vs. 503.72) with an AIC of 154.6, highlighting its good fit while suggesting potential for simplification by removing non-significant terms.

Model 1: above_avg_likes ~ 1					
Model 2: Initial Model					
	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	363	503.72			
2	354	134.60	9	369.12	< 2.2e-16 ***

Table 2: ANOVA for Initial Model

The ANOVA test proves that the addition of predictors significantly improves model fit when compared to the null model ($p < 2.2e-16$). We can also see a decrease in deviance when using our initial model.

New.Comments	Shares	New.Subscribers	Views	Watch Time
1.705780	3.738289	1.287817	3.484644	2.540871
Day Type	Shares:Day.Type	New.Comments:Day.Type	Shares:New.Subscribers	
7.007736	4.953716	5.190233	2.297604	

Table 3: VIF Table for Initial Model

To investigate multicollinearity, we chose to look at the Variance Inflation Factor values. Most of our values did not exhibit signs of multicollinearity (with VIFs close to or well below 5). The only variable of concern was Day Type, which had a VIF of 7.

	Actual	
Predicted	0	1
0	176	14
1	15	159

Table 4: Accuracy Table for Initial Model

The confusion matrix shows that the model correctly predicted 176 true negatives and 159 true positives, with 14 false negatives and 15 false positives, with an overall accuracy of 92%.

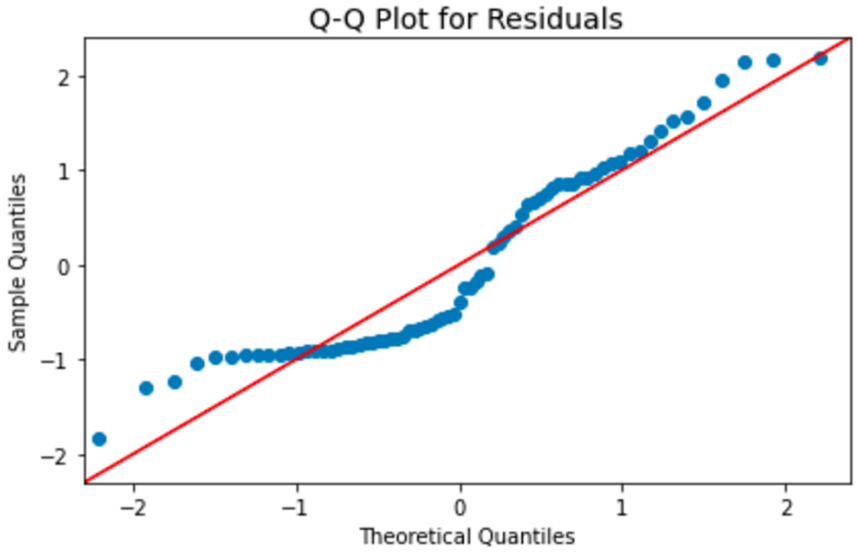


Figure 13: QQ plot for Initial Model

The Q-Q plot shows that the residuals mostly align with the red diagonal line, indicating a roughly normal distribution. However, slight deviations at the tails suggest potential minor issues with normality, which we could look into later on.

Forward Selection (Model 2)

Coefficients	Estimate	Std. Error	Z value	Pr(>|z|)
(Intercept)	-5.885e+00	7.139e-01	-8.244	< 2e-16 ***
Watch Time	3.966e-04	6.799e-05	5.833	5.43e-09 ***
Shares	9.450e-03	2.657e-03	3.557	0.000376 ***
New.Subscribers	2.388e-03	9.714e-04	2.459	0.013940 *
Null Deviance	503.72	Degrees of Freedom	363	
Residual Deviance	142.07	Degrees of Freedom	360	
AIC	150.07			
Number of Fisher Scoring iterations	8			

Table 5: Summary Table for Forward Selection Model

After using forward selection, we had a model which kept Shares, New Subscribers, and Watch time as significant predictors. This model has a lower AIC and deviances compared to our initial model.

	Odds Ratio	2.5%	97.5%
Intercept	0.00277984	0.0005801	0.0097415
Shares	1.00949476	1.0045743	1.0150725
New Subscribers	1.00239125	1.0005353	1.0044693
Watch Time	1.00039670	1.0002712	1.0005398

Table 6: Odd Ratio Table for Forward Selection

The odds ratio table shows that Shares, New Subscribers, and Watch Time have positive effects on the target outcome, with odds ratios of 1.0095, 1.0024, and 1.0004, respectively. The narrow confidence intervals indicate precise estimates, reinforcing the importance of these predictors in explaining the target variable.

Keeping Only Significant Predictors (Model 3)

Coefficients	Estimate	Std. Error	Z value	Pr(>|z|)
(Intercept)	-6.067e+00	7.444e–01	-8.150	3.63e-16 ***
Shares	1.078e-02	2.798e-03	3.855	0.000116 ***
New.Subscribers	3.178e-03	1.010e-03	3.147	0.001649 **
Watch Time	3.915e-04	6.733e-05	5.815	6.05e-09 ***
Shares: New Subscribers	-4.210e-06	1.379e-06	-3.054	0.002260 **
Null Deviance	503.72	Degrees of Freedom	363	
Residual Deviance	141.48	Degrees of Freedom	359	
AIC	151.48			
Number of Fisher Scoring iterations	9			

Table 7: Summary Table for Keeping all Significant Predictors Model

This model uses predictors deemed significant in our initial model. We can see a slightly higher AIC and lower Residual Deviance, but both change by very small amounts.

	Odds Ratio	2.5%	97.5%
Intercept	0.00231799	0.0004518	0.008928
Shares	1.01084251	1.0041116	1.016752
New Subscribers	1.00318281	0.9996529	1.005320

Watch Time	1.00039162	1.0002671	1.000532
Shares: New Subscribers	0.99999579	0.9999949	NA

Table 8: Odd Ratio Table for Keeping all Significant Predictors Model

We can once again see that Shares, New Subscribers, and Watch Time have positive effects on the response, with the interaction effect proving to be slightly less significant.

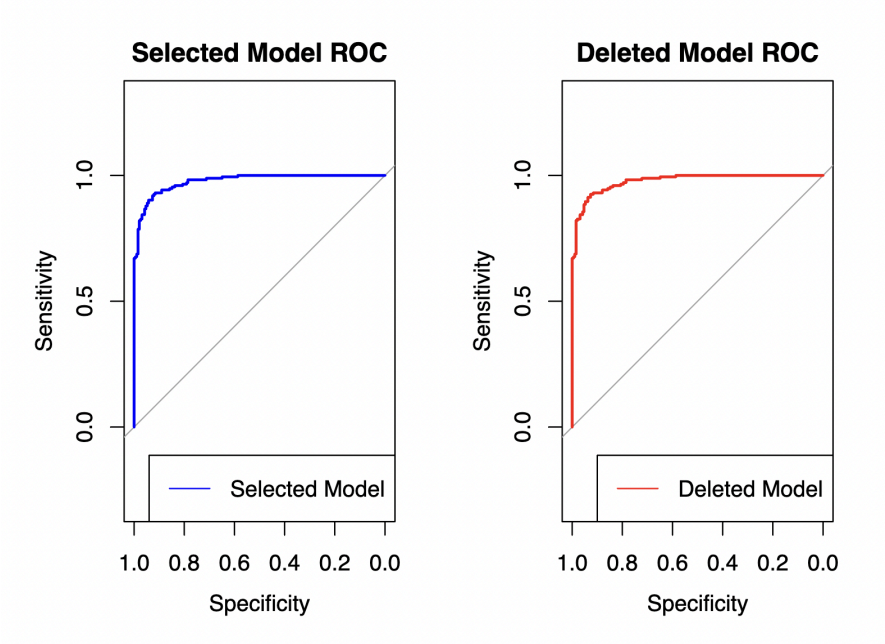


Figure 14: ROC Plots Comparing Models 2 and 3

The ROC curves of both models indicate strong classification accuracy.

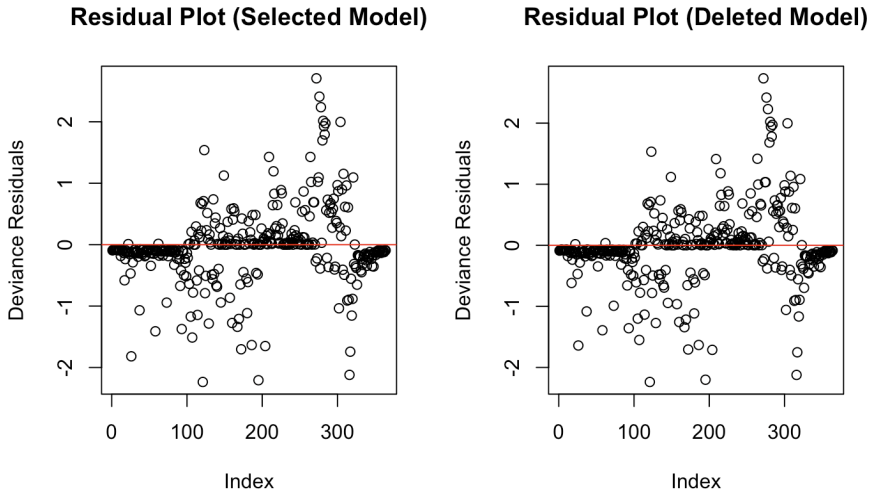


Figure 15: Residual Plots for Models 2 (left) and 3 (right)

The Residual Plots for both models look quite similar, with most points coalescing around 0. However, we can see a couple of outliers which will require further exploration.

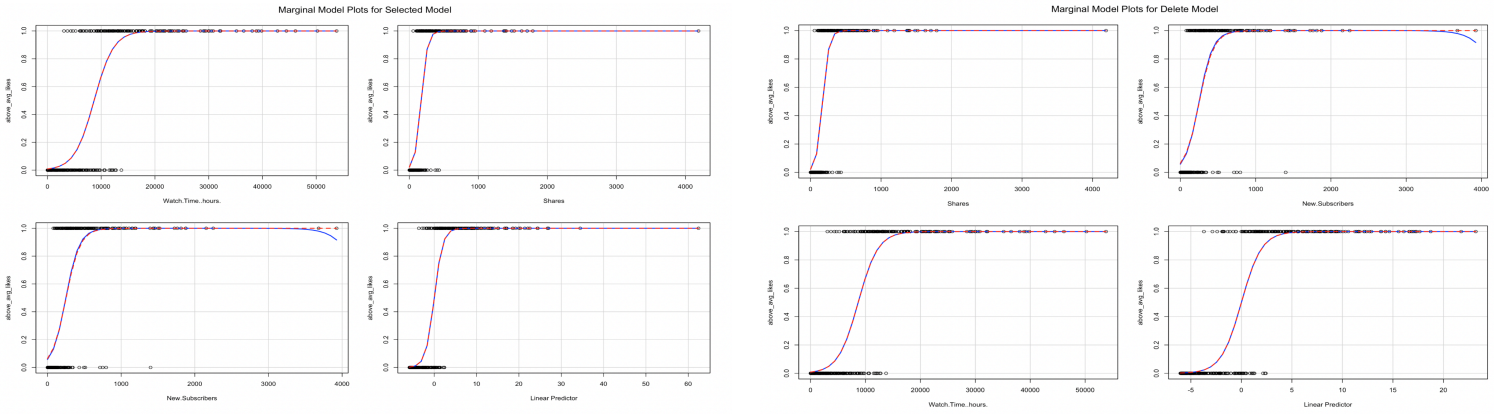


Figure 16: Marginal Model Plots for Models 2 (left) and 3 (right)

The Marginal Model Plots for both models show a strong relationship between the predictors and response, as seen in the uniformity of the S-Curves (which closely follow observed data points).

Model 1: Forward Selected Model					
Model 2: Keep Significant Model					
	Resid. Df	Resid. Dev	Df	Deviance	Pr(>Chi)
1	360	142.06			
2	359	141.47	1	0.59012	0.4424

Table 9: ANOVA for both Models

Finally, we looked at the deviances of both models using an ANOVA test. The analysis of deviance table showed no significant difference between the two models ($p = 0.4424$), indicating that the interaction effect between Shares and New Subscribers (the only additional/different predictors between both models) does not improve model fit.

Results

We explored interaction effects in the model, but none of them showed a significant impact on improving the model's performance. Additionally, we attempted log transformations on some variables; however, this approach resulted in most predictors becoming non-significant, reducing the model's effectiveness. Based on these findings, we finalized a model with three key predictors: Shares, New Subscribers, and Watch Time, as they consistently demonstrated strong significance and explanatory power.

Final Model (Model 4)

	Estimate	SD	Z value	P-value
Intercept	- 8.132	1.168	-6.964	3.31e-12 ***
Shares	0.01208	0.003370	3.586	0.000336 ***
New Subscribers	0.00317	0.001092	2.905	0.003678 **
Watch Time	0.0005672	0.00009637	5.885	3.98e-09 ***
Null Deviance	493.895	Degrees of freedom	356	
Residual Deviance	100.17	Degree of freedom	353	
AIC	108.17			
Number of Fisher Scoring Iteration	9			

Table 10: Summary Table for Final Model (Outlier Removed)

Our final model is a logistic regression model using predictors deemed significant from our initial model, with any outliers removed. The AIC (108.17 < ~150) and deviances (493.895 < ~503 and 100.17 < ~141, for Residual and Null respectively) for this model are significantly lower than models 2 and 3.

	Odds Ratio	2.5%	97.5%
Intercept	0.0002940531	0.000021111	0.002154574
Shares	1.0121556053	1.006010794	1.019343415
New Subscribers ^10	1.032213	1.011038	1.056809
Watch Time ^10	1.005688	1.003966	1.007805

Table 11: Odd Ratio for Final Model

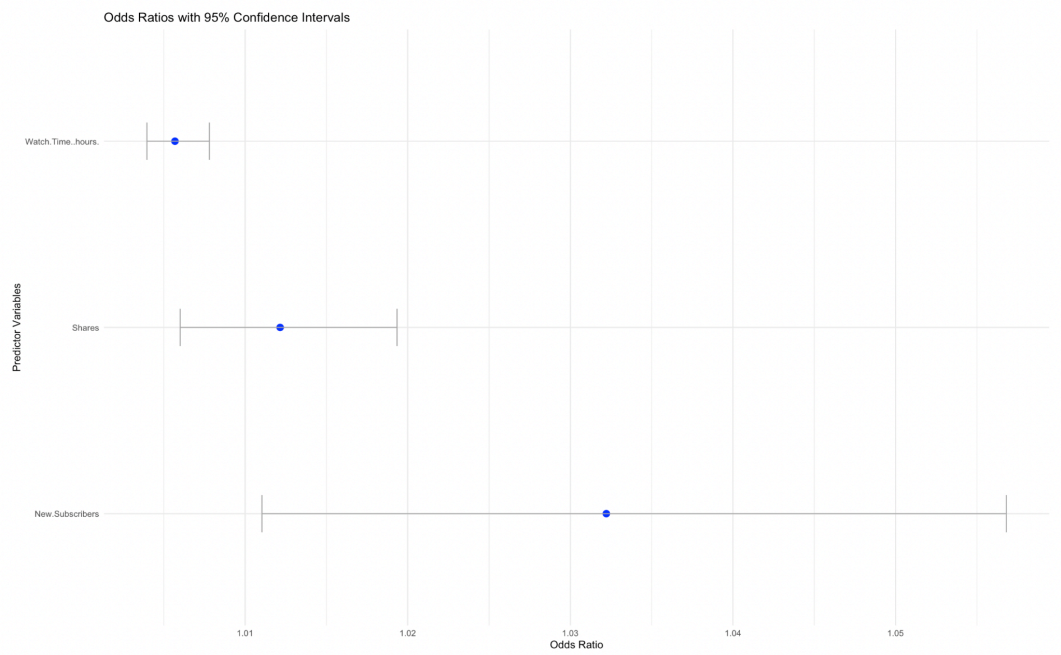


Figure 17: Odds Ratio Plot including Confidence Intervals for the Final Model

The odds ratio table and plot for the final model highlight the impact of Shares, New Subscribers (^10), and Watch Time (^10) . A unit increase in Shares increases the odds of achieving above-average likes by 1.2% (*OR*=1.012), while a 10-unit increase in New Subscribers (^10) and Watch Time (^10) increases the odds by 3.2% (*OR*=1.032) and 0.6% (*OR*=1.006), respectively. The narrow confidence intervals for all predictors suggest robust and reliable estimates, confirming their importance in the final model.

	Actual	
Predicted	0	1
0	176	9
1	12	160

Table 12: Accuracy Table for Final Model

Based on the confusion matrix, the accuracy rate of the model is 94.11%. This high accuracy indicates that the model is effective at distinguishing between categories and provides reliable predictions for the target variable.

Figure 18: Residual Plot for the Final Model

The residual plot for the refitted model after removing outliers shows a similar pattern to before, with most residuals centered around zero and no obvious systematic bias. The distribution of residuals appears slightly tighter and more uniform.

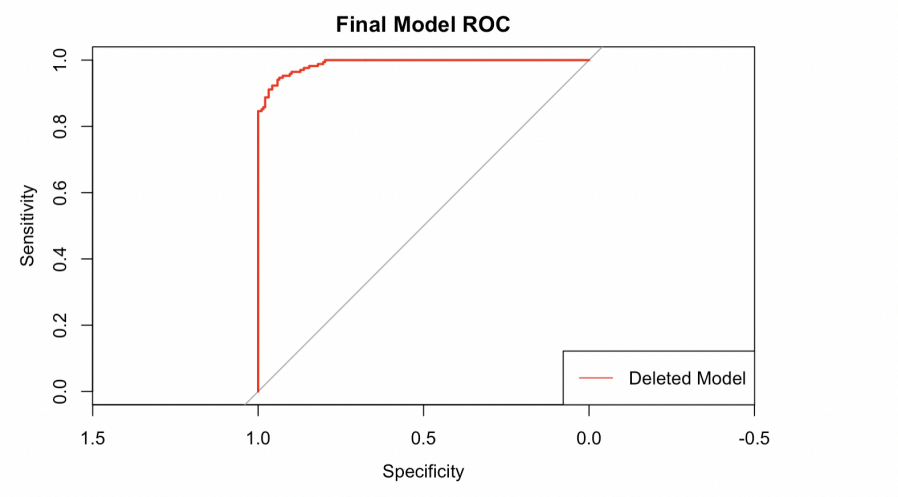


Figure 19: ROC Plot for the Final Model

The ROC curve for the Final Model (Deleted Model) demonstrates excellent classification performance, with the curve closely approaching the top-left corner. This indicates high sensitivity and specificity, confirming the model's strong predictive ability.

Limitations and Recommendations

Limitations:

1. We chose a fraction of the predictors in the original dataset, so we could have chosen completely different variables to explore.
2. We did not consider all possible interaction effects.
3. We chose to measure the “Days of the Week” predictor by weekday vs. weekend, rather than by the individual days (on the basis of having more data).

Recommendations for Future Research:

1. What factors dramatically increase creator revenue? Video length, video shares, likes, etc.
2. How does this information impact channels and channel interactions with viewers as a whole?

Conclusion

We chose model 4 as our final model because it had the most significant predictors when compared to models 2 and 3. The model's performance was demonstrated by several key metrics: the Akaike Information Criterion (AIC) for model 4 was 108.17, which was lower than both model 2 (150.07) and model 3 (151.48), indicating a better fit. The residual deviance for Model 4 was 100.17, which was also lower than Model 3 (141.48) and Model 2 (142.07), further supporting its superior fit. Additionally, the null deviance for model 4 was 493.895, which was smaller than both model 3 and model 2 (both had a null deviance of 503.72).

Based on these results, we conclude that when a YouTube video has greater shares, new subscribers, and watch time in hours, it has a higher probability of receiving above-average likes. The odds ratios for the significant predictors in Model 4 further support this conclusion:

- **Shares (Odds Ratio: 1.01217):** For each additional share, the odds of getting above-average likes increase by approximately 1.21%, assuming other variables are held constant.
- **New Subscribers^10 (Odds Ratio: 1.0322)**: Every 10 new subscribers increases the odds of receiving above-average likes by about 3.22%, keeping other variables constant.
- **Watch Time (Odds Ratio: 1.0056)**: With all other variables held constant, every 10 additional hours of watch time slightly increases the odds by approximately 0.56%.

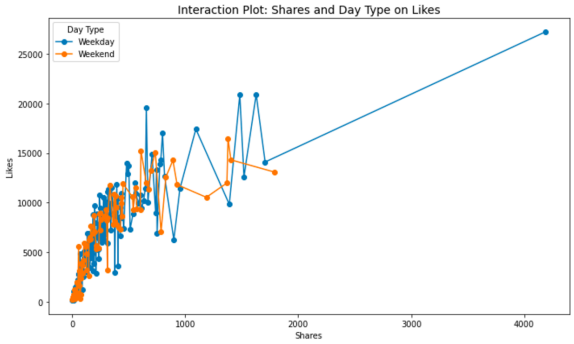
We believe content creators could greatly benefit from the insights above. This study highlights the importance of emphasizing shareability, an increase in shares significantly boosts the likelihood of receiving above average likes. Creators who wish to increase their like count can thus attempt to produce content which is more likely to be shared, or simply by communicating the importance of sharing to their viewers. The impact of New Subscribers on likes points towards the importance of attracting and retaining subscribers, perhaps pushing creators to maintain a more consistent schedule or to interact with their audience more. Creators could attempt to foster stronger connections with their viewers in order to maximize this variable.

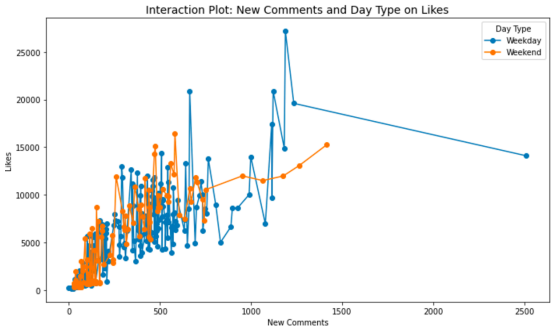
Citation:

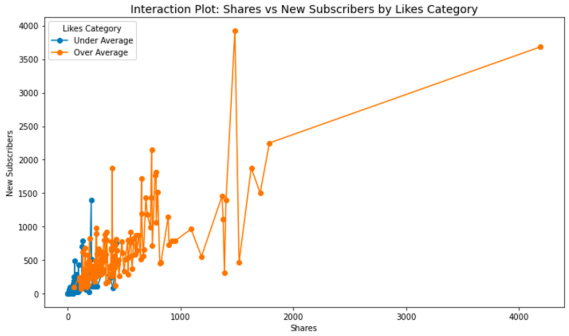
L3WY. “YouTube Channel Performance Analytics.” *Kaggle.com*, 2024,

www.kaggle.com/datasets/positivealexey/youtube-channel-performance-analytics.

Appendix







We explored interaction effects using interaction plots, including Shares and Day Type, New Comments and Day Type, and Shares and New Subscribers. While these plots provide some visual insights into potential relationships, they do not reveal meaningful or consistent patterns that significantly improve the model's interpretability or performance. Based on this observation, we decided to exclude interaction terms from the final model to maintain simplicity and focus on the main predictors.

Logged Model

Coefficients	Estimate	Std. Error	Z value	Pr(>|z|)
(Intercept)	-3.24e+01	2.949e+1	-1.117	0.264
Shares (logged)	4.053e+00	5.654e+00	0.717	0.473
New Subscribers (logged)	2.889e+00	5.250e+00	0.550	0.582
Watch Time	4.995e-04	9.760e-05	5.117	3.1e-07 ***
Shares (logged): New Subscribers (logged)	-2.869e-01	1.002e+00	-0.286	0.775
Null Deviance	493.895	Degrees of Freedom	356	
Residual Deviance	92.489	Degrees of Freedom	352	
AIC	102.49			
Number of Fisher Scoring iterations	10			

Table 13: Summary Table for Logged Model

We experimented with a logged model to address potential skewness, but most predictors, including Shares and New Subscribers, became statistically insignificant ($p>0.05$). Only Watch Time remained significant ($p=3.1\times10$^-7), indicating the transformation did not improve the

model's overall performance. As a result, we decided to revert to the original scale for better predictor importance and interpretability.