TikTok Analysis

How are views affected by followers, likes, shares, and comments?

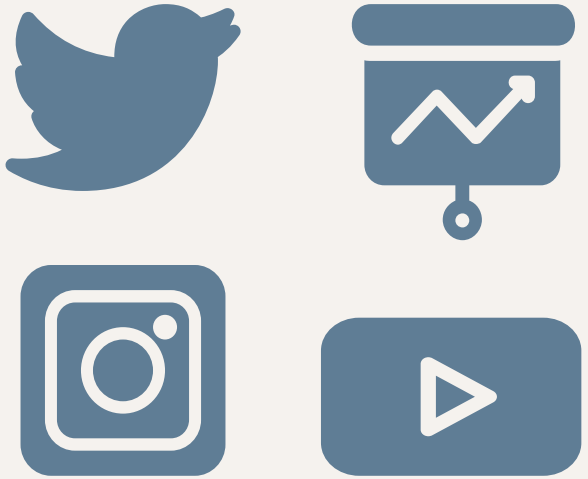
Tyler Chia, Ethan Lee, Ryan Tolone, Erica Yee

Table of contents

01 Introduction

Motivation

Learning how follower count, average likes, shares, and comments affect views can help influencers determine in which areas they need more engagement and better refine their content.



Dataset

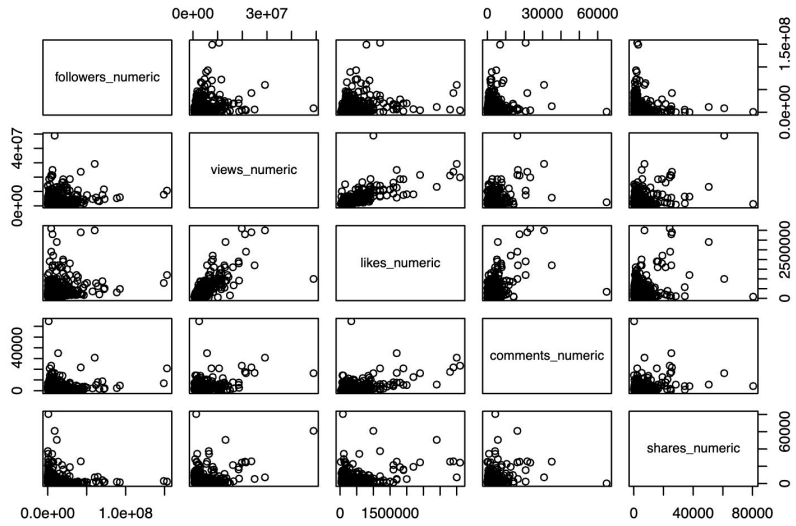
- Kaggle: Social Media Influencers in 2022

  https://www.kaggle.com/datasets/ramjasmaurya/top-1000-social-media-channels

- 1000 observations and 8 variables

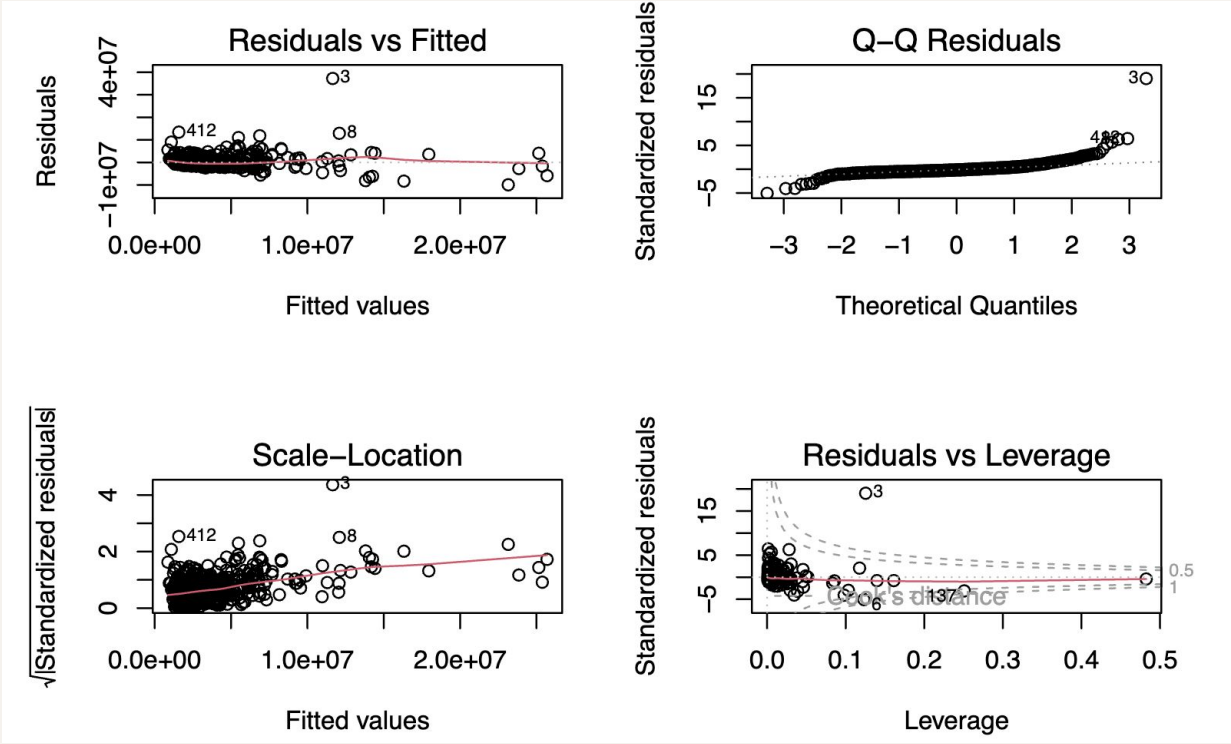
Original Model

```
pairs(tiktok_subset)
```

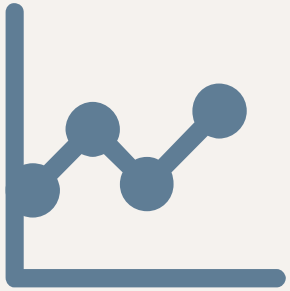



```
## Call:
## lm(formula = views_numeric ~ followers_numeric + likes_numeric +
##     comments_numeric + shares_numeric, data = tiktok_subset)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -9951230  -870340  -336551   445022 37243203
##
## Coefficients:
##                    Estimate Std. Error t value Pr(>|t|)
## (Intercept)       5.250e+05  9.955e+04   5.274 1.64e-07 ***
## followers_numeric 2.308e-02  5.707e-03   4.043 5.67e-05 ***
## likes_numeric     6.439e+00  2.374e-01  27.128  < 2e-16 ***
## comments_numeric  5.726e+00  2.300e+01   0.249    0.803
## shares_numeric    7.232e+01  1.344e+01   5.380 9.29e-08 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 2092000 on 995 degrees of freedom
## Multiple R-squared:  0.5768, Adjusted R-squared:  0.5751
## F-statistic:   339 on 4 and 995 DF,  p-value: < 2.2e-16
```

```
## Analysis of Variance Table
##
## Response: views_numeric
##                    Df     Sum Sq    Mean Sq   F value    Pr(>F)
## followers_numeric   1 6.4308e+14 6.4308e+14  146.9142 < 2.2e-16 ***
## likes_numeric       1 5.1585e+15 5.1585e+15 1178.4909 < 2.2e-16 ***
## comments_numeric    1 6.7755e+12 6.7755e+12    1.5479    0.2137
## shares_numeric      1 1.2669e+14 1.2669e+14   28.9438  9.29e-08 ***
## Residuals         995 4.3554e+15 4.3772e+12
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Residuals vs Fitted
Residuals
Fitted values
3
412
8
−1e+07
4e+07
0.0e+00
1.0e+07
2.0e+07
Q−Q Residuals
Standardized residuals
Theoretical Quantiles
30
−3
−2
−1
0
1
2
3
−5
5
15
Scale−Location
√|Standardized residuals|
Fitted values
0
2
4
Residuals vs Leverage
Standardized residuals
Leverage
137
6
Cook's distance
0.5
1
0.0
0.1
0.2
0.3
0.4
0.5

02 Methods



Handling Outliers

- Remove extreme anomalies to make more reliable model for understanding general TikTok trends
- Regression models were extremely poor when outliers and leverage points were included
- Use new cleaned dataset to create more useful model

```
residuals <- rstudent(model)
cooks_distance <- cooks.distance(model)
n <- nrow(tiktok_subset)
outliers <- which(abs(residuals) > 2 | cooks_distance > 4 / (n - 2))
cleaned_data <- tiktok_subset[-outliers, ]
```

Box-Cox Transformation

```
cleaned_data$comments_numeric[cleaned_data$comments_numeric == 0] <- 1
summary(tranxy <- powerTransform(cbind(views_numeric, followers_numeric, likes_numeric,
                                       comments_numeric, shares_numeric) ~ 1, data = cleaned_data))
```

```
## bcPower Transformations to Multinormality
##                    Est Power Rounded Pwr Wald Lwr Bnd Wald Upr Bnd
## views_numeric       -0.0803        0.00      -0.1694       0.0089
## followers_numeric    0.1860        0.19       0.1497       0.2222
## likes_numeric        0.1286        0.13       0.0515       0.2058
## comments_numeric     0.1928        0.19       0.1536       0.2319
## shares_numeric      -0.0757       -0.08      -0.1275      -0.0240
##
## Likelihood ratio test that transformation parameters are equal to 0
##  (all log transformations)
##                                     LRT df       pval
## LR test, lambda = (0 0 0 0 0) 266.0291  5 < 2.22e-16
##
## Likelihood ratio test that no transformations are needed
##                                     LRT df       pval
## LR test, lambda = (1 1 1 1 1) 4812.938  5 < 2.22e-16
```

```
cleaned_data$transformed_views <- log(cleaned_data$views_numeric)
cleaned_data$transformed_followers <- log(cleaned_data$followers_numeric)
cleaned_data$transformed_likes <- log(cleaned_data$likes_numeric)
cleaned_data$transformed_comments <- log(cleaned_data$likes_numeric)
cleaned_data$transformed_shares <- log(cleaned_data$shares_numeric)
```

- Used Box-Cox to address nonconformities to linear assumptions and create better model fit
- Opted to use log transformation for all variables since the lambdas were all close to zero

Forward Stepwise Variable Selection

- Address multicollinearity by doing variable selection

```
## Start:  AIC=-1158.15
## transformed_views ~ 1
##
##                          Df Sum of Sq    RSS     AIC
## + transformed_likes       1   122.896 156.58 -1706.0
## + transformed_comments    1   122.896 156.58 -1706.0
## + transformed_followers   1    26.303 253.17 -1250.0
## + transformed_shares      1     4.731 274.74 -1172.3
## <none>                                279.48 -1158.2
##
## Step:  AIC=-1705.95
## transformed_views ~ transformed_likes
##
##                          Df Sum of Sq    RSS     AIC
## + transformed_followers   1    4.2463 152.33 -1730.0
## + transformed_shares      1    1.3383 155.24 -1712.1
## <none>                                156.58 -1706.0
##
```

- AIC is lowest for the model that includes all predictor variables except comments

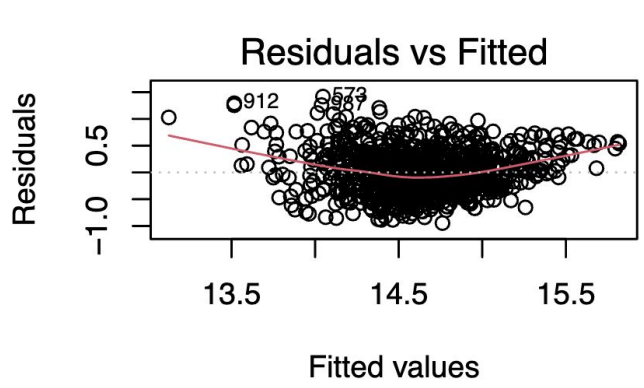
```
## transformed_views ~ transformed_likes + transformed_followers
##
##                          Df Sum of Sq    RSS     AIC
## + transformed_shares      1   0.41734 151.91 -1730.7
## <none>                                152.33 -1730.0
##
## Step:  AIC=-1730.65
## transformed_views ~ transformed_likes + transformed_followers +
##     transformed_shares
##
##        Df Sum of Sq    RSS     AIC
## <none>              151.91 -1730.7
```

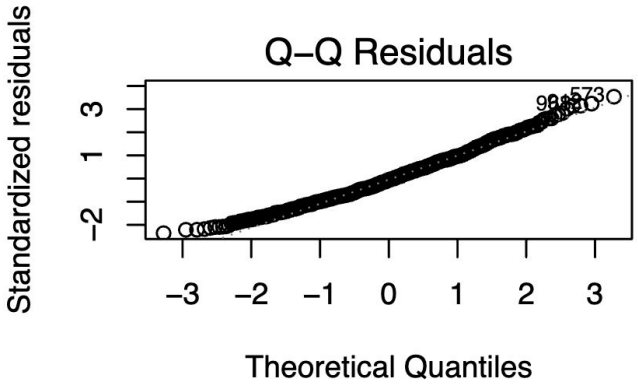
```
## Call:
## lm(formula = transformed_views ~ transformed_followers + transformed_likes,
##     data = cleaned_data)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.94614 -0.29981 -0.02219  0.26205  1.41715
##
## Coefficients:
##                       Estimate Std. Error t value Pr(>|t|)
## (Intercept)            6.46052    0.29147  22.165  < 2e-16 ***
## transformed_followers  0.05150    0.01003   5.135 3.42e-07 ***
## transformed_likes      0.59473    0.02377  25.024  < 2e-16 ***
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.4013 on 946 degrees of freedom
## Multiple R-squared:  0.4549, Adjusted R-squared:  0.4538
## F-statistic: 394.8 on 2 and 946 DF,  p-value: < 2.2e-16
```

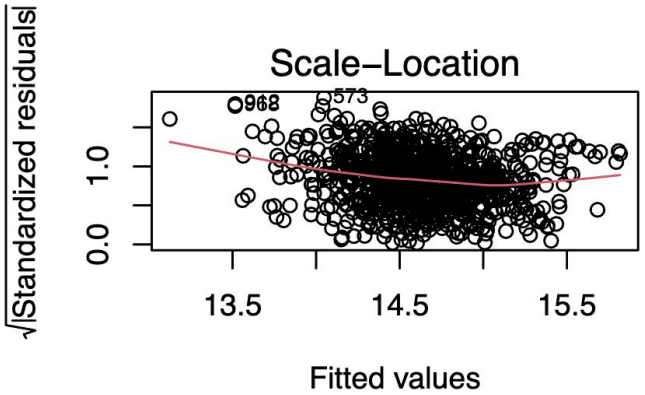
Numerical Output for Final Model

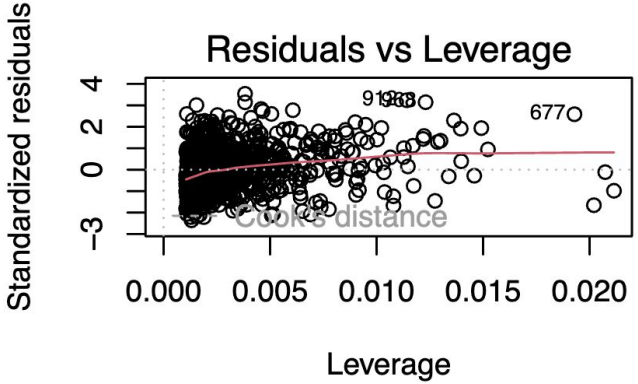
- All predictor variables are significant except “shares”
- Further investigation shows that shares don’t meaningfully contribute to the model and AIC with and without shares are nearly identical

```
## Analysis of Variance Table
##
## Response: transformed_views
##                           Df  Sum Sq Mean Sq F value    Pr(>F)
## transformed_followers      1  26.303  26.303  163.35 < 2.2e-16 ***
## transformed_likes          1 100.839 100.839  626.22 < 2.2e-16 ***
## Residuals                946 152.333   0.161
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
```


Residuals vs Fitted
Residuals
Fitted values
912
573
13.5
14.5
15.5
−1.0
0.5


Q–Q Residuals
Standardized residuals
Theoretical Quantiles
573
−3
−2
−1
0
1
2
3


Scale–Location
√|Standardized residuals|
Fitted values
573
13.5
14.5
15.5
0.0
1.0


Residuals vs Leverage
Standardized residuals
Leverage
Cook's distance
677
0.000
0.005
0.010
0.015
0.020
−3
0
2
4

03 Conclusion

Final Model Equation:

$$\ln(\text{Views}) = 6.461 + 0.052 * \ln(\text{Followers}) + 0.595 * \ln(\text{Likes})$$

- A 1% increase in followers leads to a 0.052% increase in views
- A 1% increase in average likes increases views by 0.595%

What does this mean?

- It indicates a positive relationship for followers and especially views, reinforcing the value of growing a dedicated audience and underscoring likes as the main priority for content popularity and engagement.

Because of the challenges faced while developing the model:

- Normality of errors
- Homoscedasticity
- Influence points

The model may not account for all factors, such as content quality or external trends.

- Despite challenges, the final model offers insights into the influence of likes, followers, and shares on view counts as 45.38% (R-squared value) of the variation in the view counts can be explained by the model.

- The overall model significance, as indicated by the F-statistic's p-value, is less than 2.2e-16, which is far below the alpha level of 0.05. This shows that the model is statistically significant at predicting views for tiktokers.