# **Do You Have What It Takes to Go Viral?**

## **Lesson Overview**

Duration: 60 minutes

Level: High School Statistics

Prerequisites: Basic statistics concepts, no prior R coding experience required

Materials Needed

* Computers with internet access
* Student worksheets
* TikTok .csv data file

Before Class

* Test the Web R environment to ensure the TikTok dataset loads correctly
* Review common R errors students might encounter
* Consider student grouping if sharing computers
* Ensure CSV file is accessible to all students

## **Introduction (20 minutes)**

This lesson uses TikTok data to introduce sampling distributions and the Central Limit Theorem. Students explore real social media data before formalizing statistical concepts.

In this lesson, we'll investigate what makes TikTok videos successful by analyzing real data. Can we predict which videos will go viral? Let's find out by examining TikTok engagement patterns.

Here is some TikTok data for 10 random videos taken from the TikTok platform:

| **Video** | **Category** | **Duration (sec)** | **Views** | **Likes** | **Engagement Rate (%)** | **Time Posted** | **Hashtags** |
| --- | --- | --- | --- | --- | --- | --- | --- |
| A | Dance | 15 | 105,000 | 12,600 | 12.0 | 8 PM | 4 |
| B | Sports | 45 | 28,500 | 2,280 | 8.0 | 3 PM | 6 |
| C | Comedy | 22 | 89,000 | 9,790 | 11.0 | 7 PM | 3 |
| D | Tutorial | 58 | 15,000 | 900 | 6.0 | 11 AM | 5 |
| E | Dance | 18 | 95,000 | 11,400 | 12.0 | 9 PM | 3 |
| F | Gaming | 31 | 42,000 | 2,940 | 7.0 | 4 PM | 7 |
| G | News | 25 | 68,000 | 4,760 | 7.0 | 8 AM | 4 |
| H | Comedy | 20 | 82,000 | 9,020 | 11.0 | 8 PM | 3 |
| I | Food | 35 | 31,000 | 2,170 | 7.0 | 2 PM | 5 |
| J | Music | 12 | 93,000 | 10,230 | 11.0 | 7 PM | 4 |

1. What patterns do you notice about views and engagement?

Expected student observations from the initial 10-video table:

* Shape: Right-skewed distribution
* Mean Views: ≈64,850
* SD: ≈32,640

*Teaching Tip:* Students may initially focus on absolute numbers. Guide them to notice patterns:

* 1. Evening posts (7-9 PM) consistently get more views
  2. Shorter videos (12-22 sec) tend to perform better
  3. Dance/Comedy/Music have higher engagement rates
  4. 3-4 hashtags seem optimal

1. Make three predictions about what factors lead to higher views. Explain your reasoning.

Common student predictions:

1. "Evening posts will get more views" (Data supports this)
2. "Shorter videos perform better" (True up to a point)
3. "More hashtags = more views" (Actually peaks at 3-5)

*Teaching Tip:* Let students discover that small samples (n=5) show high variability. This sets up the need for larger samples.

1. Take a random sample of 5 videos from our dataset and calculate their mean view count. Do this three times. Answers will vary based on their sample.

\*\*\*Now you will need to get the students on their computers and onto the web-r platform. They will need access to the tiktok\_data.csv file as well.

1. Now let’s take some actual TikTok data to look for trends. Please navigate to https://webr.r-wasm.org/latest/ and in the top right pane click Upload File and choose the tiktok\_data.csv file provided by your instructor. Before we analyze the data we need to load those capabilities into the platform so please copy and paste the following into the first three lines in the currently empty top left pane.

library(ggplot2)

library(dplyr)

library(tidyr)

In order for the libraries to load, you may need to look in the bottom left pane and type in a 1 to select “Yes” to have the program load them. Once you do that three times, let’s make sure the dataset is loaded and that we can read it correctly by pasting the following code underneath:

tiktok\_data <- read.csv("/tiktok\_data.csv", stringsAsFactors = TRUE)

str(tiktok\_data)

head(tiktok\_data)

summary(tiktok\_data)

As a teacher, you should run them through the pane to point out some information that these three commands gives us a window into.

1. Now let’s start creating some visuals to help us really look into this data. First let’s just look at a distribution of views:  
    ggplot(tiktok\_data, aes(x = Views)) +

geom\_histogram(bins = 20, fill = "skyblue", color = "black") +

labs(title = "Distribution of TikTok Video Views",

x = "Views",

y = "Frequency") +

theme\_minimal()

According to your data, what is the approximate range of views this random sample of videos? From about \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ views to about \_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_\_ views.

1. Okay, now let’s drill down into some categories. Go ahead and copy and paste this code over the code of the last plot you made:

ggplot(tiktok\_data, aes(x = Category, y = Views, fill = Category)) +

geom\_boxplot() +

labs(title = "Views by Content Category",

x = "Category",

y = "Views") +

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 45, hjust = 1),

legend.position = "none")

What categories seem very popular according to your dataset?

Which ones seem to have less views?

**Teaching Tips**

* Encourage students to think about why certain categories perform better
* Discuss how seasonality might affect categories
* Connect to students' own experiences with content

1. If we define the Engagement Rate as likes/views, let’s see which categories seem to actually cause a viewer to hit that heart button.

ggplot(tiktok\_data, aes(x = Category, y =

as.numeric(Engagement\_Rate), fill = Category)) +

geom\_boxplot() +

labs(title = "Engagement Rate by Content Category",

x = "Category",

y = "Engagement Rate (%)") +

theme\_minimal() +

theme(axis.text.x = element\_text(angle = 45, hjust = 1),

legend.position = "none")

Do the most viewed categories also have the highest engagement rates? Use specific

data to back up your claim.

1. Now let’s analyze the data to see if there are any trends in video duration.

# Create a scatter plot of duration vs. views

ggplot(tiktok\_data, aes(x = Duration, y = Views)) +

geom\_point(aes(color = Category), alpha = 0.7) +

geom\_smooth(method = "loess", se = TRUE, color = "black") +

labs(title = "Relationship Between Video Duration and Views",

x = "Duration (seconds)",

y = "Views") +

theme\_minimal()

# Check correlation

cor.test(tiktok\_data$Duration, tiktok\_data$Views)

What does this graph tell you about optimal video duration?

1. Now that we know how long our videos should be, let’s look to see if there is an optimal time to post a TikTok video.

# Extract hour from Time\_Posted for analysis

extract\_hour <- function(time\_str) {

# Extract components using regex

parts <- regmatches(time\_str, regexec("(\\d+):(\\d+)\\s+(AM|PM)", time\_str))[[1]]

if(length(parts) < 4) return(NA) # Return NA if pattern doesn't match

hour <- as.numeric(parts[2])

ampm <- parts[4]

# Convert to 24-hour format

if(ampm == "PM" && hour < 12) hour <- hour + 12

if(ampm == "AM" && hour == 12) hour <- 0

return(hour)

}

# Add a column with the extracted hour

tiktok\_data$hour <- sapply(tiktok\_data$Time\_Posted, extract\_hour)

# Create the plot

ggplot(tiktok\_data, aes(x = factor(hour), y = Views)) +

geom\_boxplot(fill = "skyblue") +

labs(title = "Views by Posting Hour",

x = "Hour of Day (24-hour format)",

y = "Number of Views") +

theme\_minimal()

What trends do you notice about the best times to post a video? How do those trends

seem to support people’s day to day lifestyles?

### **Common Student Misconceptions**

1. Assuming correlation equals causation
2. Not considering time zones
3. Overlooking outliers in the data

### **Sample Discussion Points**

* Why might evening hours show higher engagement?
* How might this differ on weekends vs. weekdays?
* What about different time zones?

1. Lastly, let’s see if there’s any data to support whether we should include a sound clip on our video.

ggplot(tiktok\_data, aes(x = Uses\_Trending\_Sound, y = Views,

fill = Uses\_Trending\_Sound)) +

geom\_boxplot() +

labs(title = "Impact of Trending Sounds on Views",

x = "Uses Trending Sound",

y = "Views") +

theme\_minimal() +

theme(legend.position = "none")

What does the interquartile range tell you about the difference?

What about the median?

What about the maximum?

Based on these findings, what is your recommendation about playing a trending sound clip?

1. Now let’s run a correlation analysis:

# First, ensure Engagement\_Rate is numeric

tiktok\_data$Engagement\_Rate <- as.numeric(tiktok\_data$Engagement\_Rate)

# Create correlation matrix

correlation\_data <- tiktok\_data %>%

select(Views, Likes, Shares, Comments, Duration, Hashtags, Engagement\_Rate)

# Calculate correlation matrix

correlation\_matrix <- cor(correlation\_data, use = "complete.obs")

# Install and load the corrplot package if not already available

if (!requireNamespace("corrplot", quietly = TRUE)) {

install.packages("corrplot")

}

library(corrplot)

# Create correlation plot

corrplot(correlation\_matrix,

method = "number",

type = "upper",

tl.col = "black",

col = colorRampPalette(c("darkblue", "white", "darkred"))(100),

diag = FALSE)

Based on this heatmap, what things does “views” have a strong positive correlation  
 with?

Based on this heatmap, what things does “views” have a strong negative correlation  
 with?

What are some key insights from this data that you might use if you were thinking about  
 posting your next video and wanting the most views and likes?

**Views has strong positive correlations with**:

* + Likes (0.97): Almost perfect correlation - videos with more views almost always get more likes
  + Shares (0.93): Very strong correlation - popular videos get shared more
  + Comments (0.92): Very strong correlation - more viewed videos get more comments
  + Engagement Rate (0.62): Moderate positive correlation - higher viewed videos tend to have higher engagement

**Views has a strong negative correlation with**:

* + Duration (-0.70): Longer videos tend to get fewer views - this shows that shorter videos perform better on TikTok

**Likes show similar patterns**:

* + Strong correlations with Shares (0.91) and Comments (0.90)
  + Moderate negative correlation with Duration (-0.61)
  + Strong correlation with Engagement Rate (0.77)

**Duration has negative correlations with**:

* + All engagement metrics (views, likes, shares, comments)
  + This confirms that shorter videos tend to perform better on TikTok

**Hashtags show very weak correlations**:

* + Minimal correlation with all metrics
  + This suggests the number of hashtags doesn't strongly influence performance

**Engagement Rate correlations**:

* + Strong with Likes (0.77)
  + Moderate with Views, Shares, and Comments (0.60-0.62)
  + Slight negative with Duration (-0.18)

The key insight is that video length (Duration) has a strong negative impact on performance, while the various engagement metrics (Views, Likes, Shares, Comments) are all strongly correlated with each other, suggesting that popular videos tend to perform well across all engagement dimensions.

1. What are some key insights from this lesson and data that you might use if you were thinking about posting your next video and wanting the most views and likes?

## **Exit Ticket**

Using what we've learned about sampling distributions:

1. Why do different samples of 30 videos give us different means?

Random sampling variation

Each sample captures different videos

Natural variation in population

1. What would happen to our dot plot of sample means if we used samples of 100 videos instead?

Dot plot would be more compact

Less variation between samples

More consistent estimates

SE reduces by factor of √(100/30)

1. How could a content creator use this information to improve their TikTok strategy?

Use larger datasets for strategy

Account for natural variation

Focus on controllable factors

Test across multiple samples

## **Homework Problems**

1. If the mean views for dance videos is 50,000 with σ = 15,000, what's the probability a sample of 30 videos will have a mean above 55,000?

Problem: P(x̄ > 55,000) when μ = 50,000, σ = 15,000, n = 30 Solution:

SE = σ/√n = 15,000/√30 = 2,739

z = (55,000 - 50,000)/2,739 = 1.83

P(z > 1.83) = 0.0336

Answer: About 3.36% chance

1. A content creator's videos have a mean engagement rate of 8% with σ = 2%. If they post 40 videos, what's the probability their average engagement will be between 7.5% and 8.5%?

Problem: P(7.5% < x̄ < 8.5%) when μ = 8%, σ = 2%, n = 40 Solution:

SE = 2%/√40 = 0.316%

z₁ = (7.5% - 8%)/0.316% = -1.58

z₂ = (8.5% - 8%)/0.316% = 1.58

P(-1.58 < z < 1.58) = 0.886

Answer: About 88.6% chance

1. Using the Central Limit Theorem, explain why taking a larger sample of videos gives us a more reliable estimate of true TikTok success rates.

CLT Explanation:

SE = σ/√n shows variability decreases with larger n

Class data demonstrates this directly

Connect to student samples: n=5 vs n=30

## **Connection to AP Statistics Standards**

* Sampling Distributions
* The Central Limit Theorem
* Probability and Normal Distributions
* Statistical Inference

## **Common Student Misconceptions**

1. "Bigger sample = population parameters"
   * Clarify: Better estimate ≠ exact value
2. "More hashtags always better"
   * Show: Diminishing returns after 5
3. "Viral videos prove/disprove patterns"
   * Discuss: Outliers vs typical performance

## **Extension Activities**

For advanced students:

1. Calculate confidence intervals for viral thresholds
2. Analyze interaction effects between factors
3. Design A/B testing experiments

## **Common Technical Issues & Solutions**

1. Plot not showing
   * Solution: Run all cells in order
   * Check for matplotlib inline magic
2. Input errors in final challenge
   * Solution: Guide students through error messages
   * Remind about valid input ranges
3. Data generation issues
   * Solution: Restart kernel and run all
   * Check for library imports

## **Assessment Strategies**

1. Completion of workbook
2. Quality of analysis in written responses
3. Participation in discussions
4. Understanding of key concepts:
   * Data visualization
   * Pattern recognition
   * Statistical relationships