Project Proposal

ggteam - Will Chen, Katelyn Cai, Hannah Choi, Weston Slayton

```
library(tidyverse)
library(tidymodels)
# add other packages as needed

# add code to load data
```

Introduction

TikTok now has over 1 billion users globally, and over 150 million Americans, making it one of the fastest growing social platforms in the world. As it has risen to prominence, so has its ubiquitous algorithm, which is said to generally account for account factors (likes and comments) and video information (captions, sounds, hashtags). An internal TikTok document contained by the New York Times explained the algorithm in simplistic terms: "Plike X Vlike + Pcomment X Vcomment + Eplaytime X Vplaytime + Pplay X Vplay." Essentially, likes, comments and playtime, as well as an indication that the video has been played. Given, that TikTok has been heavily criticized alongside other platforms for declining youth mental health outcomes and rising hate due to the addictive nature of its explore page, we decided to look at TikTok's data ourselves and look at what drives video views (video length, likes, shares, comments, number of hashtags, and followers). Our hypothesis is that while likes, shares, comments, number of hashtags, and followers drive up video view count, video length count will drive that down.

Data description

The dataset comes from the 'top_users_vids.csv' file (under folder 'Trending Videos Data Collection') of the Github repository found at: https://github.com/ivantran96/TikTok_famous/tree/main. The data was originally collected as part of the DataResolutions's Data Blog project exploring Tiktok's demographics and trending video analytics.

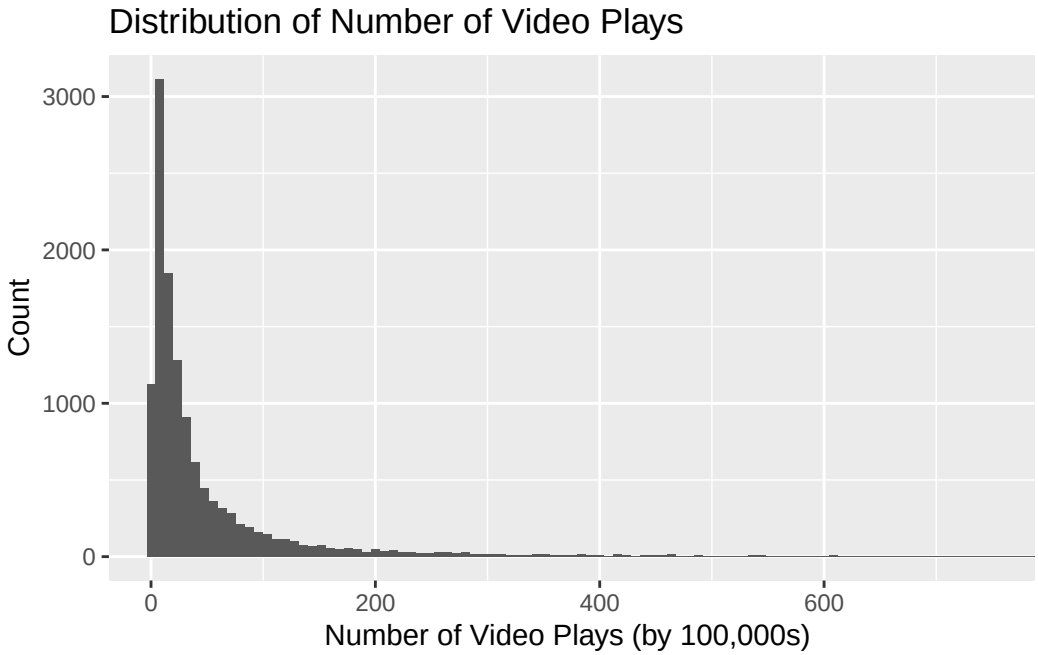
The original data curators collected the data using David Teather's open-source Unofficial Tiktok API (found at https://github.com/davidteather/TikTok-Api), which uses Python to scrape Tiktok data and fetch the most trending videos, specific user information, and much more. Using the list of top Tiktokers, the curators compiled a list of users with the getSuggestedUsersbyIDCrawler api method, which used the top TikTokers and collected the suggested users. Using the byUsername method, they collected video data of the 25 most recent posts of each user from the top TikTokers and the suggested list. The curators also used the API's bySound method to collect videos using some of the most famous songs on TikTok to get an idea of how the choice of music can impact the potential of a video to become a trending video.

The dataset has 13 columns and 12,559 rows. The columns cover important metrics for trending videos such as video length, hashtags used, songs/sounds used, and number of likes, shares, comments, plays, and followers (and their total number of likes and videos). There are also less relevant variables such as creator username, create time, video ID which we will not use in our analysis. Variables id, create_time, video_length, n_likes, n_shares, n_comments, n_plays, n_followers, n_total_likes, and n_total_vids are numerical while the others are categorical.

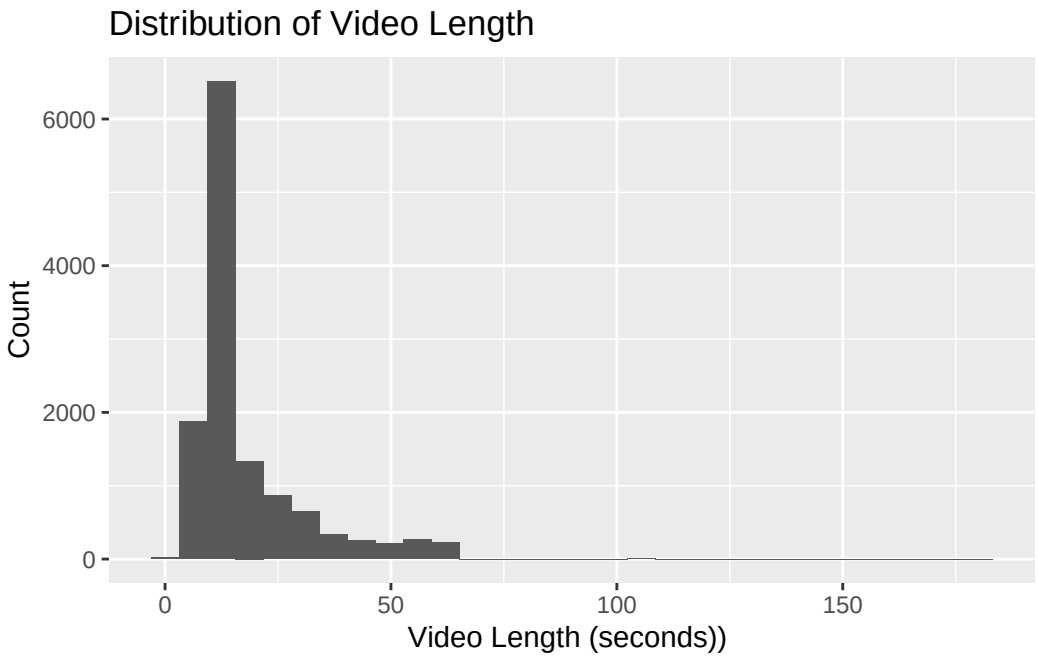
Initial exploratory data analysis

```
tiktok <- read.csv("data/top_users_vids.csv")
```

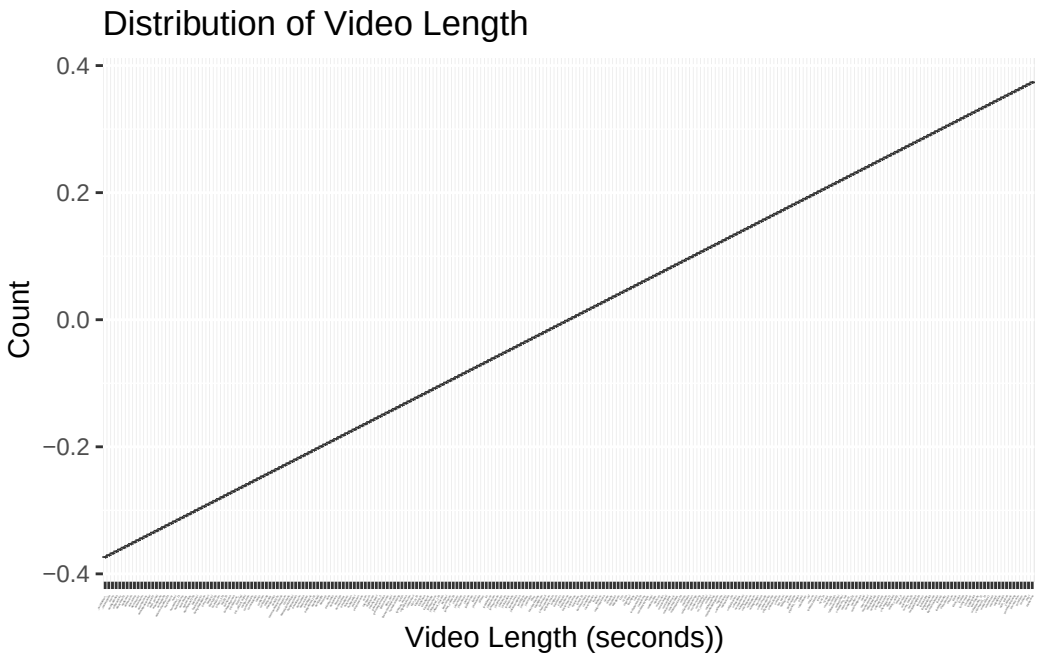
```
tiktok |>
  ggplot(aes(x = n_plays / 100000)
         ) +
  labs(x = "Number of Video Plays (by 100,000s) ", y = "Count", title = "Distribution of N
  geom_histogram(binwidth = 8) +
  coord_cartesian(xlim = c(0, 750))
```




```
ggplot(data = tiktok, aes(x = video_length)) +
  geom_histogram() +
  labs(x = "Video Length (seconds)) ", y = "Count", title = "Distribution of Video Length"
```




```
ggplot(data = tiktok, aes(x = user_name)) +
  geom_boxplot() +
  labs(x = "Video Length (seconds)) ", y = "Count", title = "Distribution of Video Length"
 theme(axis.text.x = element_text(angle = 60, hjust = 1, size = 1))
```



Analysis approach

...

Data dictionary

The data dictionary can be found here [Update the link and remove this note!]