Project Proposal

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

```r
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.

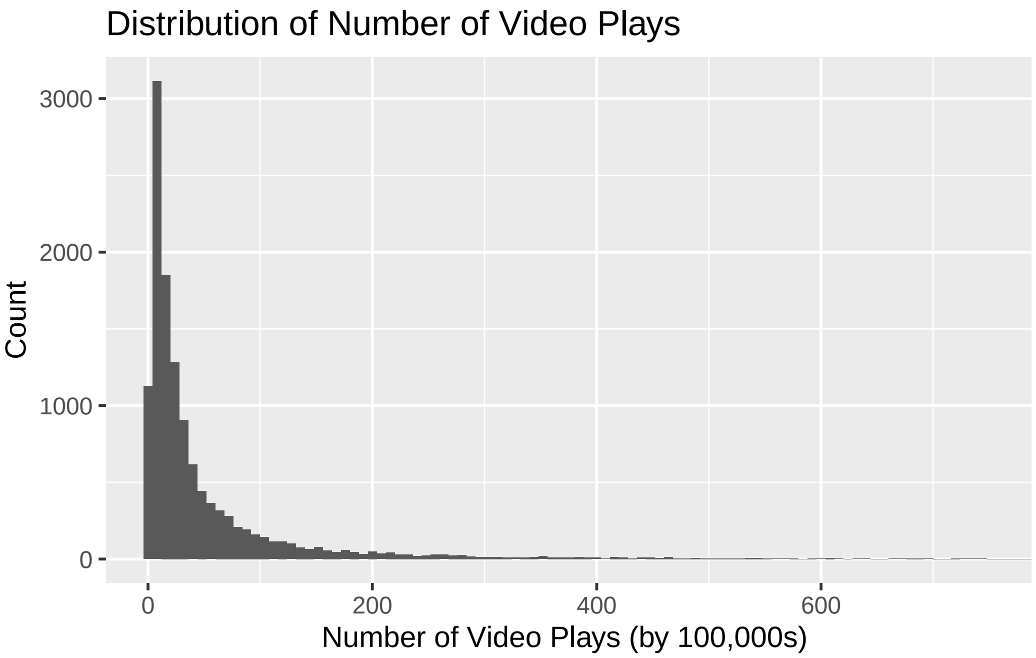
Data description

…

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))
```



Analysis approach

...

Data dictionary

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