# Executive Summary: Statistical Testing Results

TikTok Claims Classification Project

## Project Overview

The TikTok data team aims to build a machine learning model to help classify claims and opinions in user submissions. Initially, they will conduct a hypothesis test to investigate the connection between the verified status of a user's videos and the number of views each video receives.

## Key Insights

- The analysis shows that there is a difference in number of views between TikTok videos posted by verified accounts and TikTok videos posted by unverified accounts.
- As a result, these findings suggest there might be fundamental behavioral differences between these two groups of accounts: verified and unverified.
- It would be interesting to investigate the root cause of this behavioral difference. For example, consider:
    - Do unverified accounts tend to post more engaging videos? Is that engaging content a claim or opinion?
    - Or, are unverified accounts associated with spam bots that help inflate view counts?

## Details

The TikTok data team analyzed the connection between account verification (verified_status) and video view counts (video_view_count). They approached this by comparing the average video view count for each group of account types (verified and unverified) in the sample data. The results showed that unverified accounts had an average of 265,663 views, whereas verified accounts had an average of 91,439 views.

Additionally, the team conducted a two-sample hypothesis test to further investigate the relationship between verified_status and video_view_count. Consistent with the mean value findings, this statistical test suggested that any observed differences in the sample data were likely due to actual differences in the underlying population means.

## Next Steps

We recommend proceeding with building a regression model to analyze the "verified_status" feature. This regression model can provide insights into the behavior of verified users in this dataset. The findings from this regression model can then be used to inform the development of a claim classification model, which will be created afterwards. By analyzing the verified user group first, the team can leverage this context to improve the performance and interpretability of the subsequent claim classification model.