Executive Summary: Statistical Testing Results

TikTok Claims Classification Project

Project Overview

The TikTok data team is working on creating a machine learning model to help categorize claims from user submissions. In this phase of the project, it is essential to examine, explore, tidy up, and organize the data before proceeding with the development of the model.

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 considered the relationship between verified_status and video_view_count.

As part of the analysis, the mean values of video_view_count were examined for each group based on verified_status in the sample data. The results indicated that the majority of accounts were unverified, with 265,663 unverified accounts compared to 91,439 verified accounts.

```
verified_status
not verified    265663.785339
verified         91439.164167
Name: video_view_count, dtype: float64
```

The second approach involved a two-sample hypothesis test. Consistent with the initial findings from the mean values, this statistical analysis demonstrates that any observed difference in the sample data is attributed to a genuine difference in the corresponding population means.

Next Steps

The team recommends advancing to the next stage by constructing a **regression model** focusing on the verified status. A regression model for verified_status would facilitate the analysis of user behavior within this subset of verified users. Subsequently, the insights gained from this context can be taken into account when interpreting the results from a claim classification model to be developed later.