Executive Summary

Hypothesis Testing

ISSUE / PROBLEM

Our project management officer is interested in the statistical difference between verified and unverified accounts. We were instructed to conduct a hypothesis test on the correlation of video_view_count and verified_status.

RESPONSE

We will use python to conduct a hypothesis test. The null hypothesis will claim there is no significant correlation between verified_status and video_view_count, and the alternative hypothesis will claim there is. We will set our significance level to .05.

IMPACT

The hypothesis test will allow us to understand whether verified_status is an important variable for sorting. If it is the case that view counts are much larger for tiktoks from one type of account, we can make our sorting process more efficient by disproportionately targeting those.

We start by calculating the mean value of video_view_count for both verified_status values.

```
verified_status
not verified    265663.785339
verified         91439.164167
```

There is a dramatic difference in the mean of video_view_count when we group by each verified_status value. This likely indicates a relationship between the variables, but we will conduct a hypothesis test to quantify that claim.

First we created two new variables consisting of the video_view_count column but with one filtered for verified accounts and the other not. Then we used the two sample t-test to compare the significance of the differences in mean view counts given the verified status of the account.

```
not_verified = data[data["verified_status"] == "not verified"]["video_view_count"]
verified = data[data["verified_status"] == "verified"]["video_view_count"]

stats.ttest_ind(a=not_verified, b=verified, equal_var=False)
```

```
Ttest_indResult(statistic=25.499441780633777, pvalue=2.6088823687177823e-120)
```

As we expected the test gave us an extreme affirmation of our claim that verified_status and video_view_count are correlated.

KEY INSIGHTS

From our first calculation we found that the mean of the video_view_count for unverified accounts was 265,664 and for verified accounts 91,439. This gives us strong evidence that these variables would be correlated.

Since $2.61 * 10^{-120}$ is much smaller than .05, we reject the null hypothesis and conclude there is a significant relationship between the verified status of our tiktok author and its view count. Thus we should consider emphasizing tiktoks from unverified accounts to our algorithm so that it prioritizes tiktoks with high view counts. If the content being shared is problematic enough to be censored, a variable indicating high view counts would allow more efficient prioritization.