Executive Summary

Exploratory Data Analysis

ISSUE / PROBLEM

To produce a machine learning algorithm that efficiently and correctly sorts claims from opinions we must conduct a thorough investigation of our data so that we can avoid problems later in this project.

RESPONSE

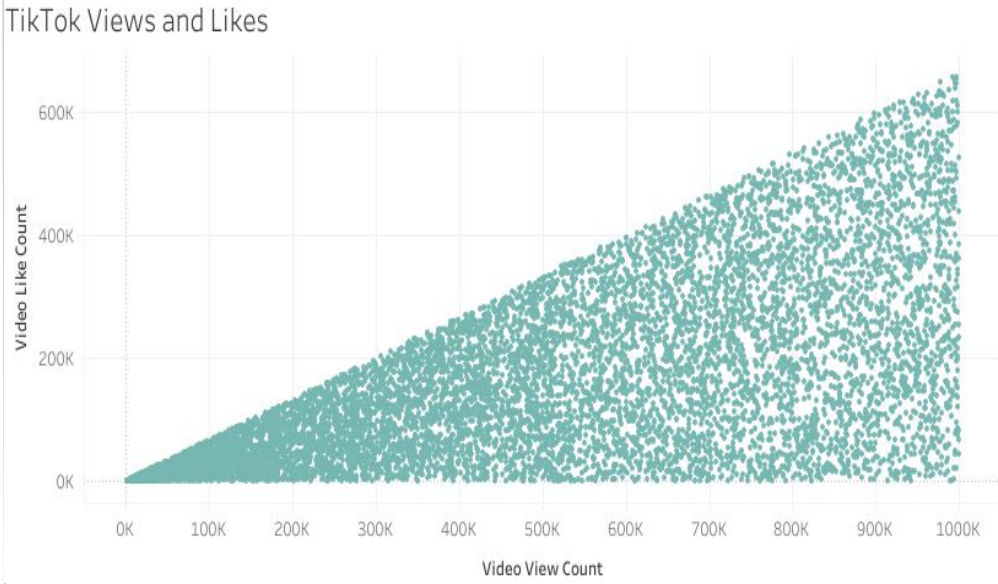
In conducting exploratory data analysis, the team filtered out some variables about engagement that seemed related to the task of the algorithm: view, like, and comment counts. Distributions of the variable types were analyzed using histograms, and several visualizations were created with Tableau to better understand the data.

IMPACT

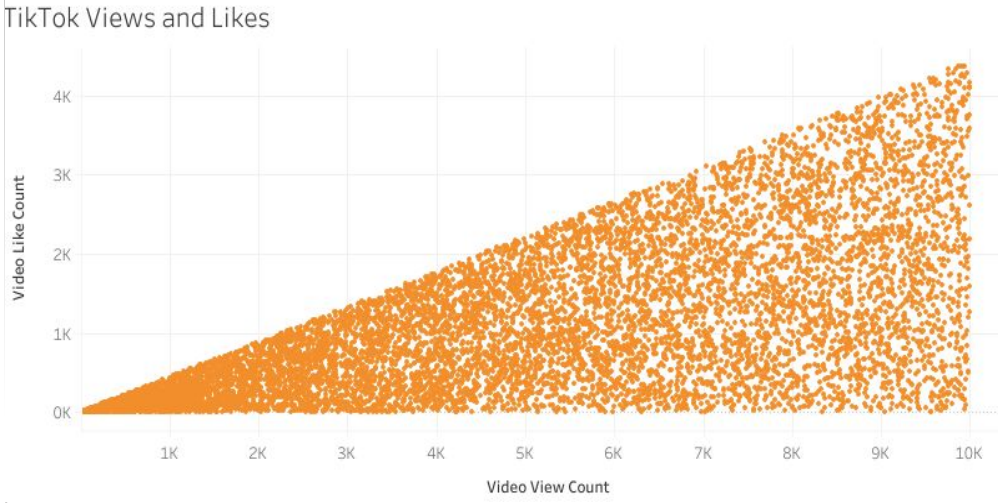
The exploratory data analysis could indicate that the classification model will need to handle null values should there be too many; it also could tell us about variable correlations that will allow us to write a more accurate and efficient algorithm.

Below are scatter plots from Tableau relating the view counts of opinion/claim tiktoks against their like counts.

Claims:



Opinions:



KEY INSIGHTS

The exploratory data analysis revealed the prevalence of many null values that could be problematic. Ideally we would be able to ask the data organizers about the implications of these missing values, but without further knowledge about the data, analysis will need adjusted to account for the missing data.

Both opinions and claims have a very consistent linear correlation between view count and like count, specifically about .65 likes/view for claims and .4 likes/view for opinions. We conclude that the disproportionate engagement with claim tiktoks will need accounted for in our final machine learning algorithm.