Exploratory Data Analysis of TikTok Claims Data

Executive summary report for TikTok

Project Overview

The TikTok data team is developing a machine learning model for classifying claims made in videos submitted to the platform. It is now time to begin the process of exploratory data analysis (EDA).

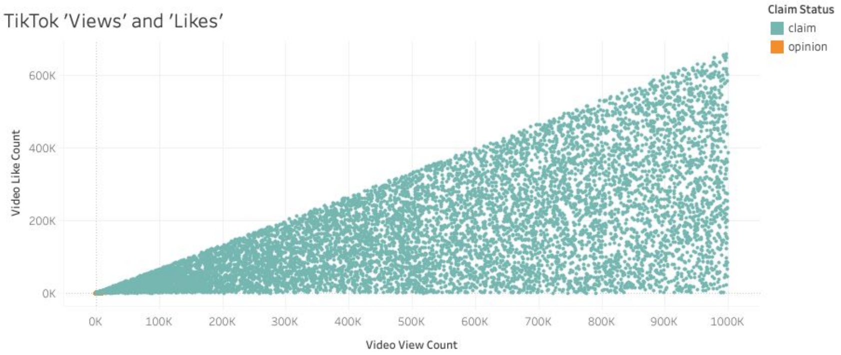
Key Insights

The Problem: After running initial exploratory data analysis (EDA) on the video claims data provided by TikTok, the values for the count variables are not normally distributed. They are heavily skewed to the right. It is good practice to get a sense of just how many data points could be considered outliers.

Proposed solution: Determine the whiskers of a column's boxplot by calculating the interquartile range (IQR). Set the lower whisker by taking the maximum between the column's minimum value and Q1 - 1.5 * IQR. Set the upper whisker by taking the minimum between the column's maximum value and Q3 + 1.5 * IQR.

Details

As a result of the conducted exploratory data analysis, the TikTok data team considered Video Like Count and Video View Count as key variables for depicting a video as a claim or as an opinion. The provided scatter plot shows the relationship between the two variables. This scatter plot was created in Tableau to enhance the provided visualization.



Alt Text: Scatter Plot displaying TikTok data plotting variables for Like Count and View Count by Claim Status

Next Steps

- Determine any unusual data points that could pose a problem for future analysis in classifying claims.
- Determine the variables that have the largest impact on claim classification.
- Filter down to consider the most relevant variables for running regression, statistical analysis, and parameter tuning.