Machine Learning Models for TikTok: Analyzing Claims and Opinions in Content

TikTok Project: Exploratory Data Analysis (EDA)

Objective

In this phase of the TikTok project, the goal is to conduct **Exploratory Data Analysis (EDA)** to better understand the dataset, uncover underlying patterns, and identify any issues that need to be addressed before building the claims classification model. We aim to visualize relationships between key variables, assess data distribution, and evaluate the overall data quality.

Phase Objectives

Data Exploration:

- Loaded dataset into pandas DataFrame, checked for missing values and data types.
- Generated summary statistics and visualized distributions (histograms, boxplots).
- Analyzed relationships between variables using scatter plots, pair plots, and correlation heatmaps.
- Detected outliers in critical variables and identified patterns, particularly between claim status and video metrics.

Visualizations:

- **Claim vs Opinion Count**: Created a bar chart comparison.
- **Boxplots**: For key variables to detect outliers.
- **Correlation Heatmap**: Examined relationships between continuous variables.
- **Stacked Bar Charts**: Visualized claims vs opinions for different video metrics in Tableau.

Tableau Dashboards:

- Created accessible, easy-to-interpret visualizations for non-technical stakeholders.

Executive Summary:

- Summarized key findings from EDA, highlighting data quality issues and insights.
- Provided recommendations for data cleaning or transformation before model building.

Machine Learning Models for TikTok: Analyzing Claims and Opinions in Content

Jupyter NoteBook:

Task 1. Imports, links, and loading

```
# Import packages for data manipulation
import pandas as pd
import numpy as np

# Import packages for data visualization
import matplotlib.pyplot as plt
import seaborn as sns
```

```
# Load dataset into dataframe
data = pd.read_csv("tiktok_dataset.csv")
```

Task 2a. Data exploration and cleaning

- I plan to begin my data exploration by utilizing methods such as .head(), .size, .shape, .info(), and .describe().

```
# Display and examine the first few rows of the dataframe
data.head()
```

	#	claim_status	video_id	video_duration_sec	video_transcription_text	verified_status	author_ban_status	video_view_count	video_like_count	video_share_count	video_download_count	video_comment_count
0	1	claim	7017666017	59	someone shared with me that drone deliveries a...	not verified	under review	343296.0	19425.0	241.0	1.0	0.0
1	2	claim	4014381136	32	someone shared with me that there are more mic...	not verified	active	140877.0	77355.0	19034.0	1161.0	684.0
2	3	claim	9859838091	31	someone shared with me that american industria...	not verified	active	902185.0	97690.0	2858.0	833.0	329.0
3	4	claim	1866847991	25	someone shared with me that the metro of st. p...	not verified	active	437506.0	239954.0	34812.0	1234.0	584.0
4	5	claim	7105231098	19	someone shared with me that the number of busi...	not verified	active	56167.0	34987.0	4110.0	547.0	152.0

```
# Get the size of the data
data.size
```

```
232584
```

```
# Get the shape of the data
data.shape
```

```
(19382, 12)
```

```
# Get basic information about the data
data.info()
```

```
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 19382 entries, 0 to 19381
Data columns (total 12 columns):
 #   Column                    Non-Null Count  Dtype  
---  ------                    --------------  -----  
 0   #                         19382 non-null  int64  
 1   claim_status              19084 non-null  object 
 2   video_id                  19382 non-null  int64  
 3   video_duration_sec        19382 non-null  int64  
 4   video_transcription_text  19084 non-null  object 
 5   verified_status           19382 non-null  object 
 6   author_ban_status         19382 non-null  object 
 7   video_view_count          19084 non-null  float64
 8   video_like_count          19084 non-null  float64
 9   video_share_count         19084 non-null  float64
 10  video_download_count      19084 non-null  float64
 11  video_comment_count       19084 non-null  float64
dtypes: float64(5), int64(3), object(4)
```

Machine Learning Models for TikTok: Analyzing Claims and Opinions in Content

```
# Generate a table of descriptive statistics
data.describe()
```

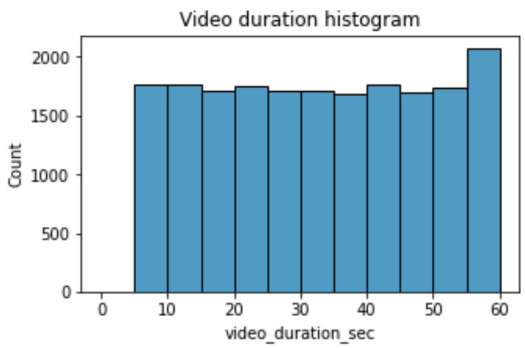
	#	video_id	video_duration_sec	video_view_count	video_like_count	video_share_count	video_download_count	video_comment_count
count	19382.000000	1.938200e+04	19382.000000	19084.000000	19084.000000	19084.000000	19084.000000	19084.000000
mean	9691.500000	5.627454e+09	32.421732	254708.558688	84304.636030	16735.248323	1049.429627	349.312146
std	5595.245794	2.536440e+09	16.229967	322893.280814	133420.546814	32036.174350	2004.299894	799.638865
min	1.000000	1.234959e+09	5.000000	20.000000	0.000000	0.000000	0.000000	0.000000
25%	4846.250000	3.430417e+09	18.000000	4942.500000	810.750000	115.000000	7.000000	1.000000
50%	9691.500000	5.618664e+09	32.000000	9954.500000	3403.500000	717.000000	46.000000	9.000000
75%	14536.750000	7.843960e+09	47.000000	504327.000000	125020.000000	18222.000000	1156.250000	292.000000
max	19382.000000	9.999873e+09	60.000000	999817.000000	657830.000000	256130.000000	14994.000000	9599.000000

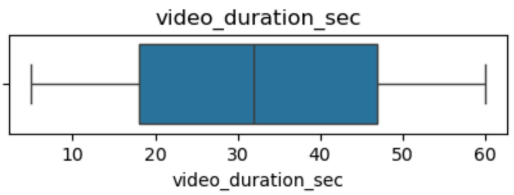
Task 2b. Select visualization type(s)

- *The visualizations I find most helpful for understanding the distribution of the data are **box plots** and **histograms**. These help in analyzing the spread and identifying patterns or outliers. Visualizing the data distribution also plays a key role in deciding the next steps in data analysis, like selecting suitable modeling techniques.*

Task 3. Build visualizations

1-Video Duration Per Second

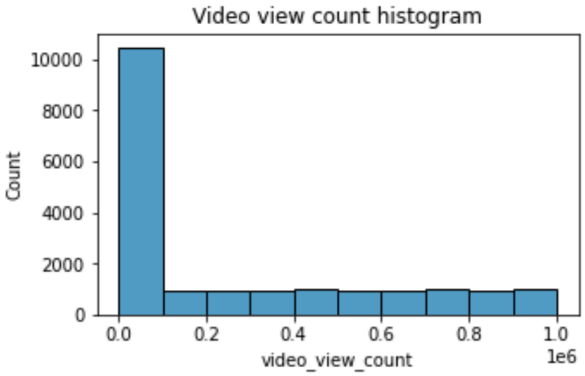


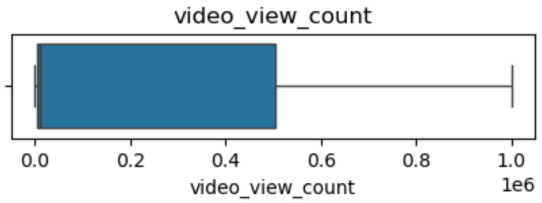


- All videos are 5-60 seconds in length, and the distribution is uniform

Machine Learning Models for TikTok: Analyzing Claims and Opinions in Content

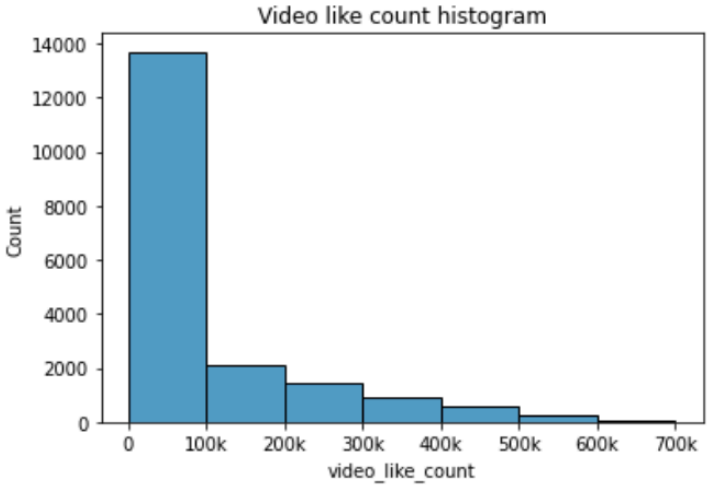
2-Video View Count

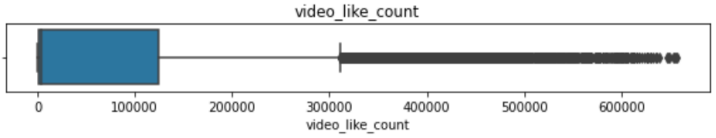




- This variable has a very uneven distribution, with more than half the videos receiving fewer than 100,000 views. Distribution of view counts > 100,000 views is uniform.

3-Video Like Count

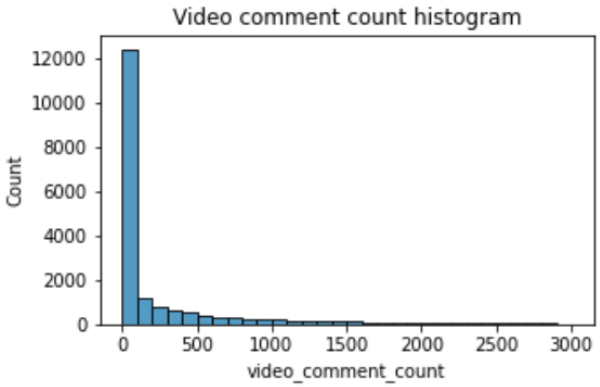


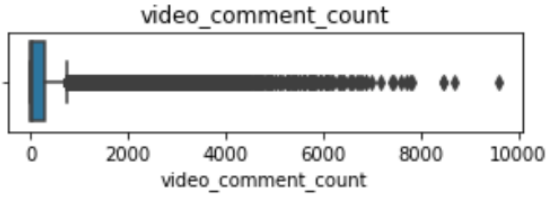


- Similar to view count, there are far more videos with < 100,000 likes than there are videos with more. However, in this case, there is more of a taper, as the data skews right, with many videos at the upper extremity of like count.

Machine Learning Models for TikTok: Analyzing Claims and Opinions in Content

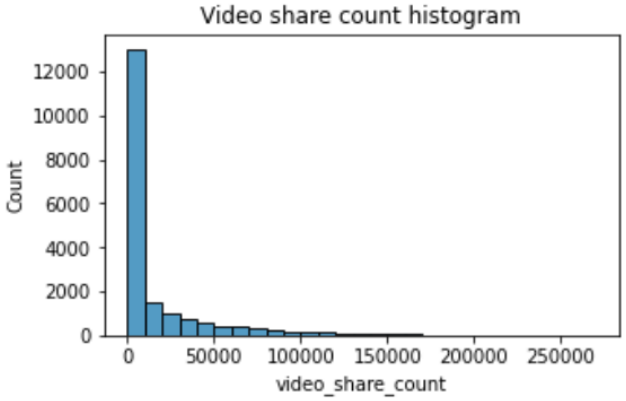
4-Video Comment Count

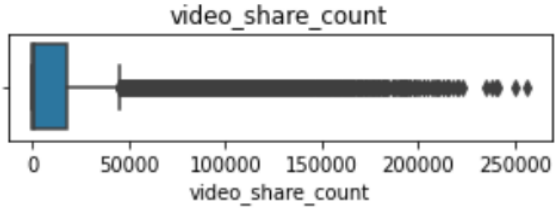




- Again, the vast majority of videos are grouped at the bottom of the range of values for video comment count. Most videos have fewer than 100 comments. The distribution is very right-skewed.

5-Video Share Count

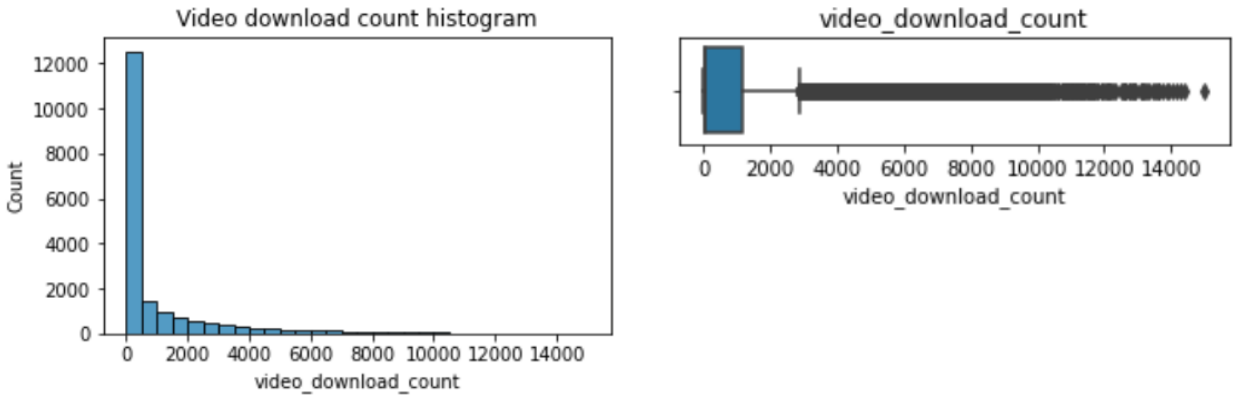




- The overwhelming majority of videos had fewer than 10,000 shares. The distribution is very skewed to the right.

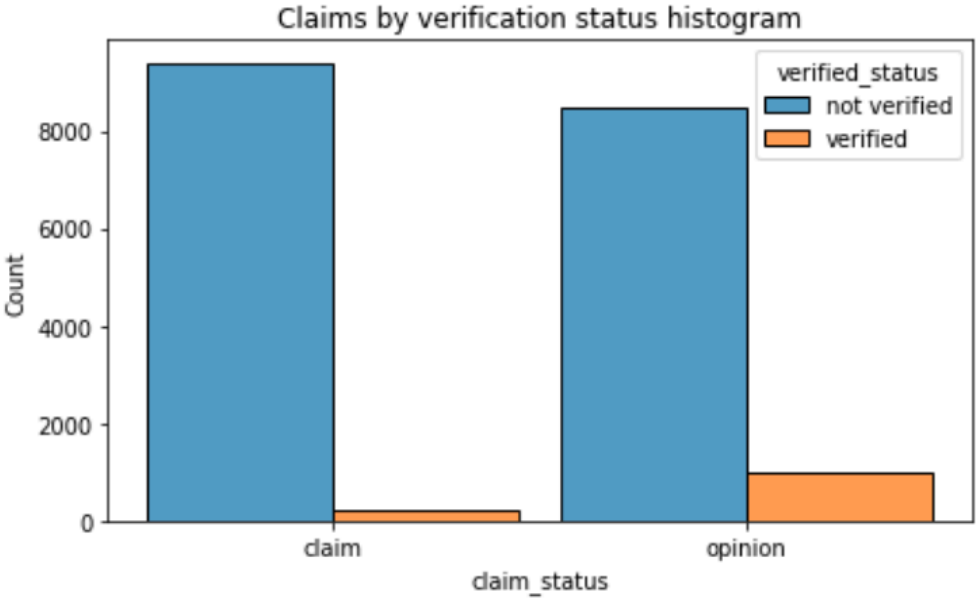
Machine Learning Models for TikTok: Analyzing Claims and Opinions in Content

6-Video Download Count



- The majority of videos were downloaded fewer than 500 times, but some were downloaded over 12,000 times. Again, the data is very skewed to the right.

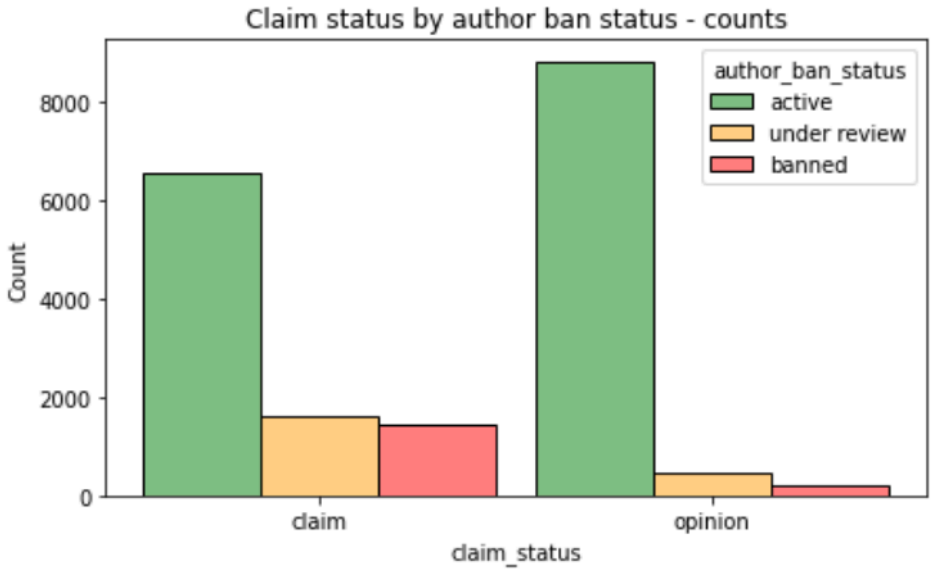
7-Claim status by verification status



- There are far fewer verified users than unverified users, but if a user is verified, they are much more likely to post opinions.

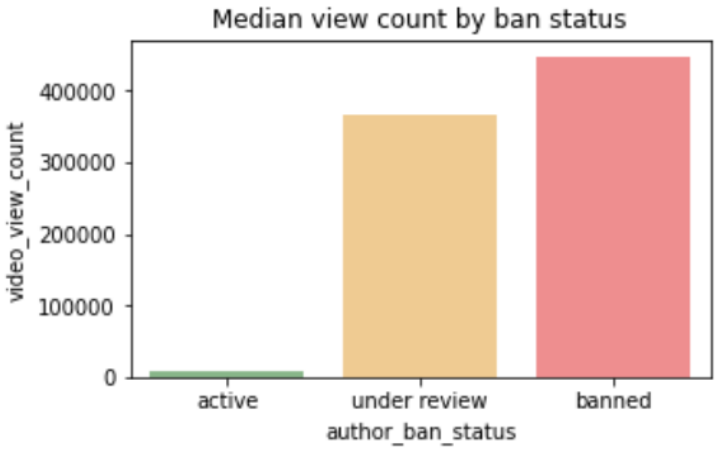
Machine Learning Models for TikTok: Analyzing Claims and Opinions in Content

8-Claim status by author ban status



- For both claims and opinions, there are many more active authors than banned authors or authors under review; however, the proportion of active authors is far greater for opinion videos than for claim videos. Again, it seems that authors who post claim videos are more likely to come under review and/or get banned.

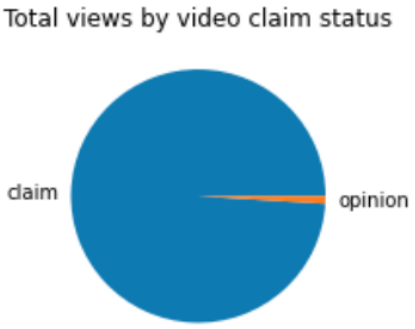
9-Median view counts by ban status



- The median view counts for non-active authors are many times greater than the median view count for active authors. Since you know that non-active authors are more likely to post claims, and that videos by non-active authors get far more views on aggregate than videos by active authors, then video_view_count might be a good indicator of claim status.

Machine Learning Models for TikTok: Analyzing Claims and Opinions in Content

10-Total views by claim status

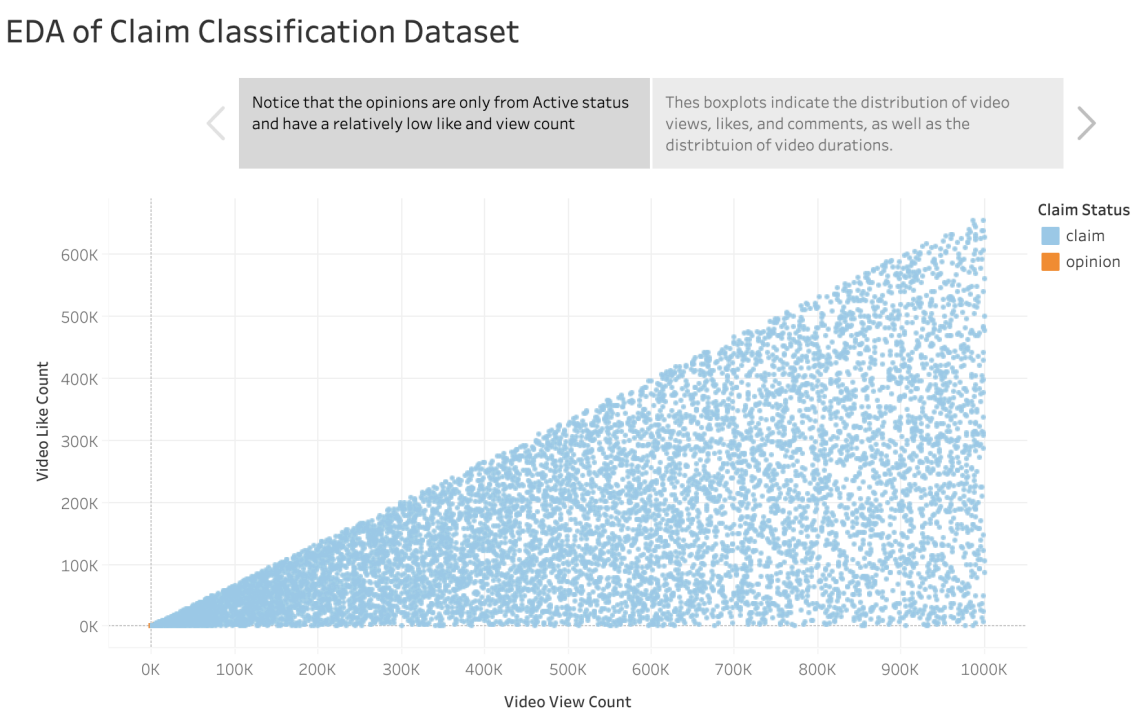


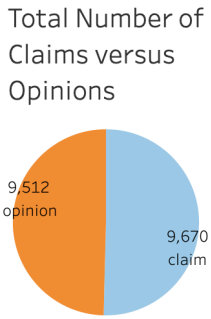
claim	501555.0
opinion	4953.0

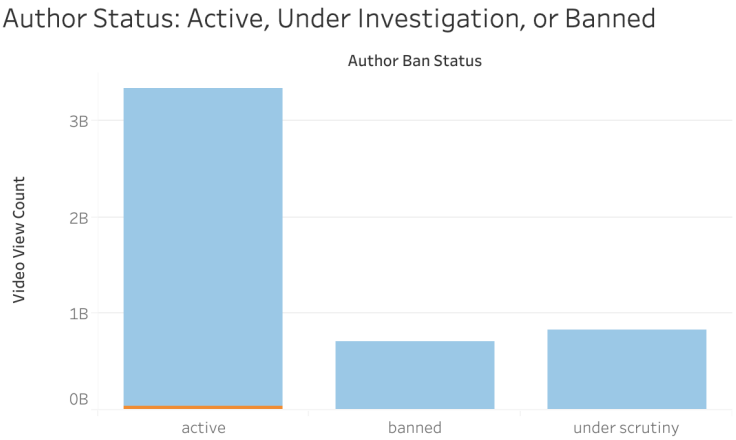
- The overall view count is dominated by claim videos even though there are roughly the same number of each video in the dataset.

Machine Learning Models for TikTok: Analyzing Claims and Opinions in Content

Tableau Dashboards:

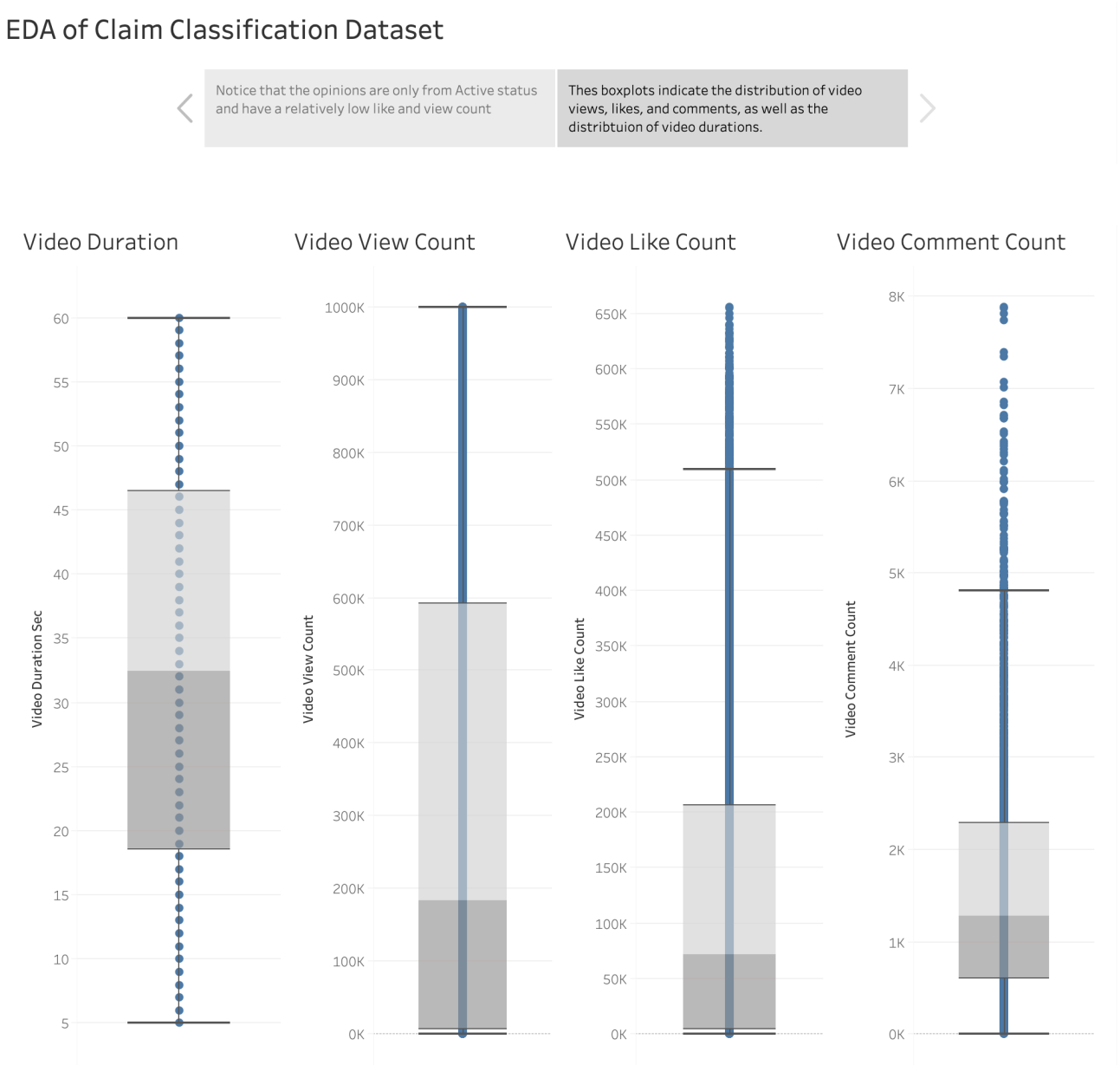






Machine Learning Models for TikTok: Analyzing Claims and Opinions in Content

Executive Summary:



Machine Learning Models for TikTok: Analyzing Claims and Opinions in Content

Tableau Dashboards:

TikTok Claims Classification Project

Exploratory Data Analysis (EDA) - Executive Summary

ISSUE / PROBLEM

The TikTok data team seeks to develop a machine learning model to assist in the classification of claims for user submissions. In this part of the project, the data needs to be analyzed, explored, cleaned, and structured prior to any model building.

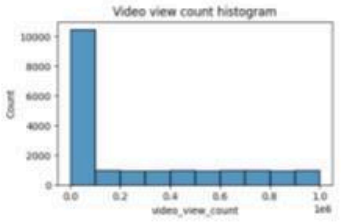
RESPONSE

The TikTok data team conducted exploratory data analysis at this stage. The purpose of the exploratory data analysis was to understand the impact that videos have on TikTok users. To do so, the TikTok data team analyzed variables that would showcase user engagement: view, like, and comment count.

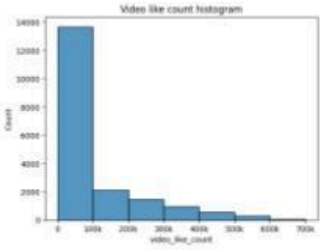
IMPACT

According to the findings from the exploratory data analysis, the future claim classification model will need to account for null values and imbalance in opinion video counts by incorporating them into the model parameters.

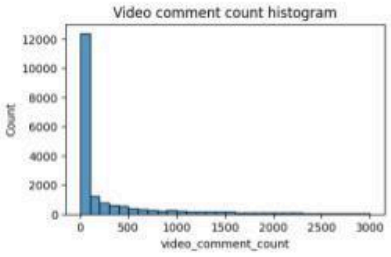
A key component of this project's exploratory data analysis involves visualizing the data. As illustrated in the following histograms, it is clear that the vast majority of videos are grouped at the bottom of the range of values for three variables that showcase TikTok users (video viewers') engagement with the videos included in this dataset.



The view count variable has a very uneven distribution, with more than half the videos receiving fewer than 100,000 views. Distribution of view counts > 100,000 views is uniform.



Similar to view count, there are far more videos with < 100,000 likes than there are videos with more.



Again, the vast majority of videos are grouped at the bottom of the range of values for video comment count. Most videos have fewer than 100 comments. The distribution is very right-skewed.

KEY INSIGHTS

The exploratory data analysis conducted from TikTok's data team revealed many considerations for the classification model, including missing values, "claims" to "opinions" balance, and overall distribution of data variables. The two key insights from this analysis were:

Null values
Over 200 null values were found in 7 different columns. As a result, future modeling should consider the null values to avoid making insights that would assume complete data. Further analysis is necessary to investigate the reason for these null values, and their

Skewed data distribution
Video view and like counts are all concentrated on low end of 1,000 for opinions. Therefore, the data distribution is right-skewed, which will inform the models and model types that will be built.