CS 621: Foundations of Data Analytics
Ball State University
Spring 2024
By Salsabil Chowdhury Tory

1. Introduction

In the era of digital content, platforms like TikTok have become a significant part of our daily lives. With the surge in user-generated content, the need for effective moderation has become paramount. Users on TikTok have the ability to report videos and comments that contain claims, leading to a large volume of user reports that need to be reviewed. This presents a challenge due to the sheer volume and the need for timely moderation.

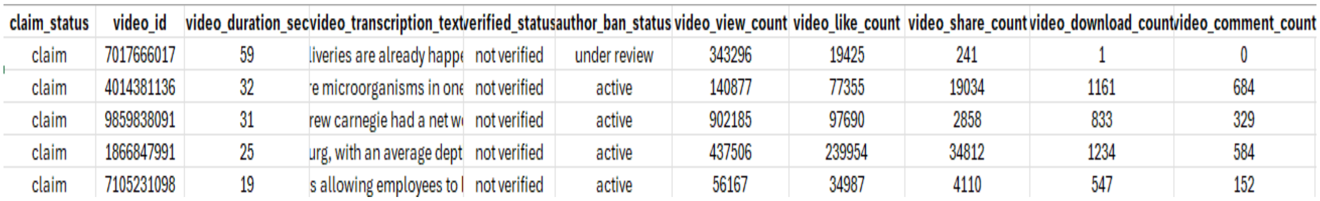
To address this, TikTok is embarking on the development of a predictive model that can determine whether a video contains a claim or an opinion. The objective of this model is to streamline the moderation process by reducing the backlog of user reports and prioritizing them more efficiently. A successful prediction model would not only enhance the user experience but also ensure a safer and more reliable platform for content sharing.

This project, as part of the Foundation of Data Analysis course, aims to demonstrate the application of data analysis skills acquired through the course. We will be working with the tiktok_dataset.csv data, applying various stages of the data science pipeline, including data ingestion, data engineering, analytics computation, and visualization. The end goal is to provide insights that could aid in the development of the predictive model for TikTok.

Through this project, we hope to highlight the practical application of data analysis techniques and contribute to the ongoing efforts in making digital platforms safer and more user-friendly. Let us embark on this data-driven journey together!

2. Data Source (5 hours)

- Data Source Description: The tiktik_dataset.csv is a comprehensive dataset that contains various metrics and information related to TikTok videos. It has been generated from user reports on TikTok, which identify content that needs to be reviewed by moderators. The dataset contains 19,382 entries and 12 columns, each representing a different attribute of the videos.
- Here is a small sample of the dataset:



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
claim	7017666017	59	liveries are already happe	not verified	under review	343296	19425	241	1	0
claim	4014381136	32	re microorganisms in one	not verified	active	140877	77355	19034	1161	684
claim	9859838091	31	rew carnegie had a net w	not verified	active	902185	97690	2858	833	329
claim	1866847991	25	urg, with an average dept	not verified	active	437506	239954	34812	1234	584
claim	7105231098	19	s allowing employees to l	not verified	active	56167	34987	4110	547	152

The dataset has 19,382 records, each representing a unique video on TikTok. The columns in the dataset are the variables that represent various attributes of the videos:

1. claim status: Represents whether the video contains a claim.
2. Video id: This is a unique identifier for each video.
3. Video duration sec: Represents the duration of the video in seconds.
4. Video transcription text: Contains the transcribed text from the video.
5. Verified status: Indicates whether the video is verified or not.
6. Author ban status: Represents the ban status of the author of the video.
7. Video view count: Represents the number of views the video has received.
8. Video like count: Represents the number of likes the video has received.
9. Video share count: Represents the number of times the video has been shared.
10. Video download count: Represents the number of times the video has been downloaded.
11. Video comment count: Represents the number of comments on the video.

This data is highly relevant to our Data Analysis project as it provides a wealth of information that can be used to develop statistical analysis, a predictive model to determine whether a video contains a claim or offers an opinion. This can help TikTok reduce the accumulation of user reports and prioritize them more efficiently.

Data Science Pipeline

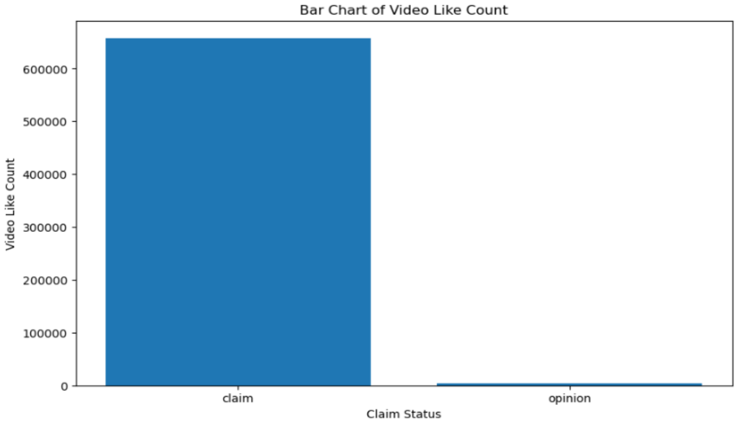
- **Data Ingestion(15 hours)**:

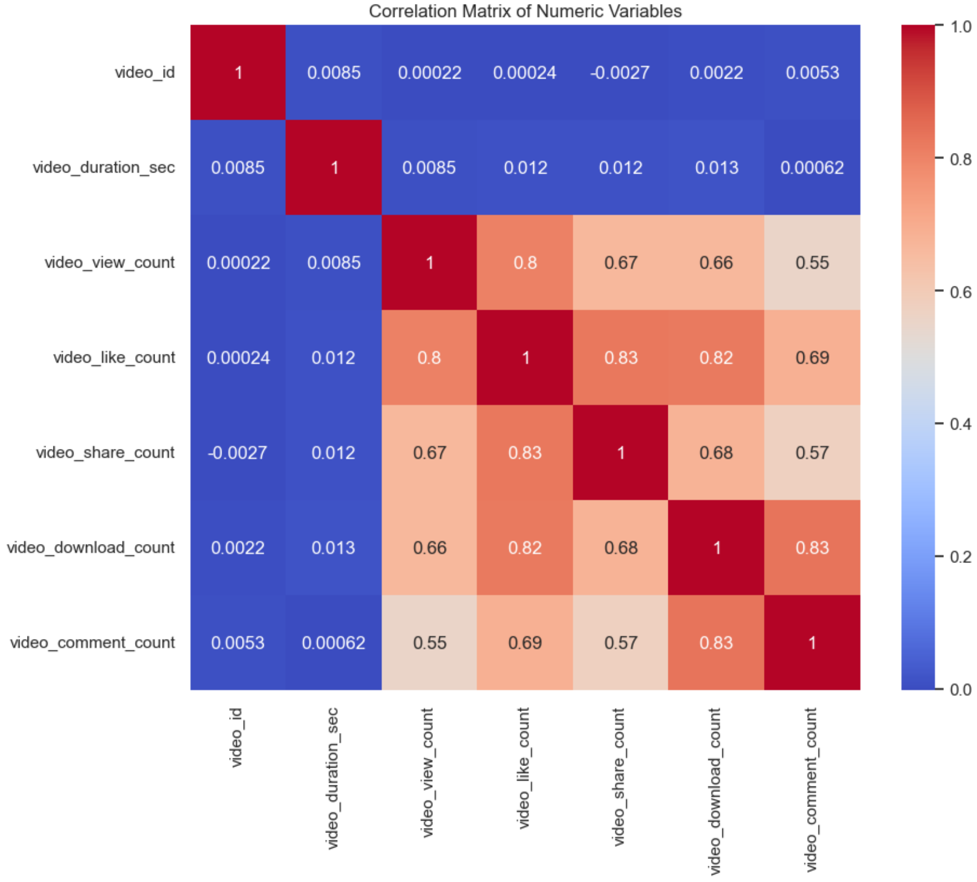
  Step1: Importing necessary libraries. We have imported pandas, matplotlib, sklearn, seaborn and NumPy.

  Step2: Loading the CSV file. We have used pandas' **read_csv** function to load the CSV file into a Data Frame, which is a 2-dimensional labeled data structure in pandas.

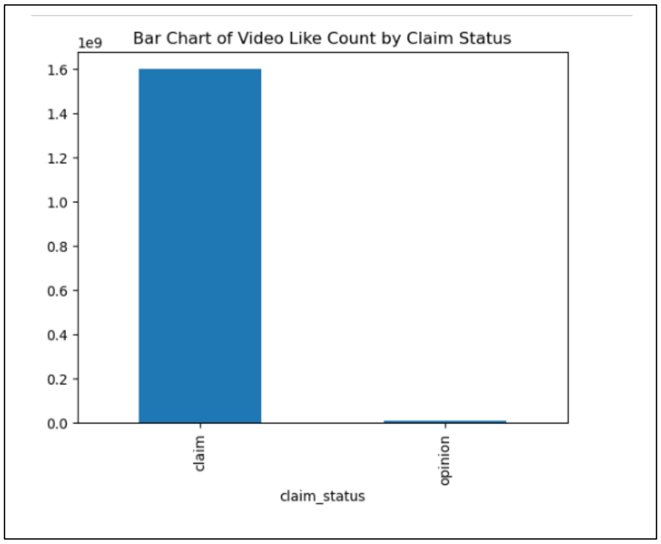
  Step3: After loading the data, we can start exploring it. We have used df.head() to display the first five rows of the dataframe. We have used df.shape() to show the number of rows and columns in the dataset. Then df.info() was used to get a summary of the dataframe, including the name of the columns and their datatypes (integer, Boolean, categorical or decimals). The df.describe() process provides the descriptive statistics of the dataframe such as mean, median, count, std, min max etc.

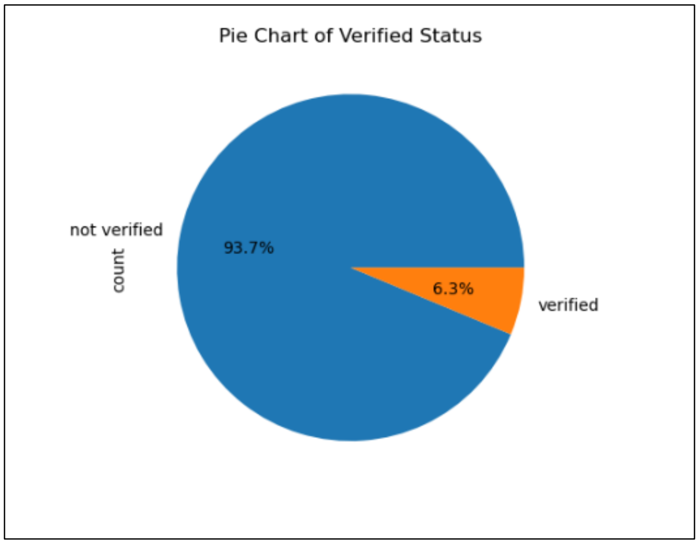
  Step 4: After that we conducted some visualizations. We have created multiple plots to visualize the various variables in the dataset.

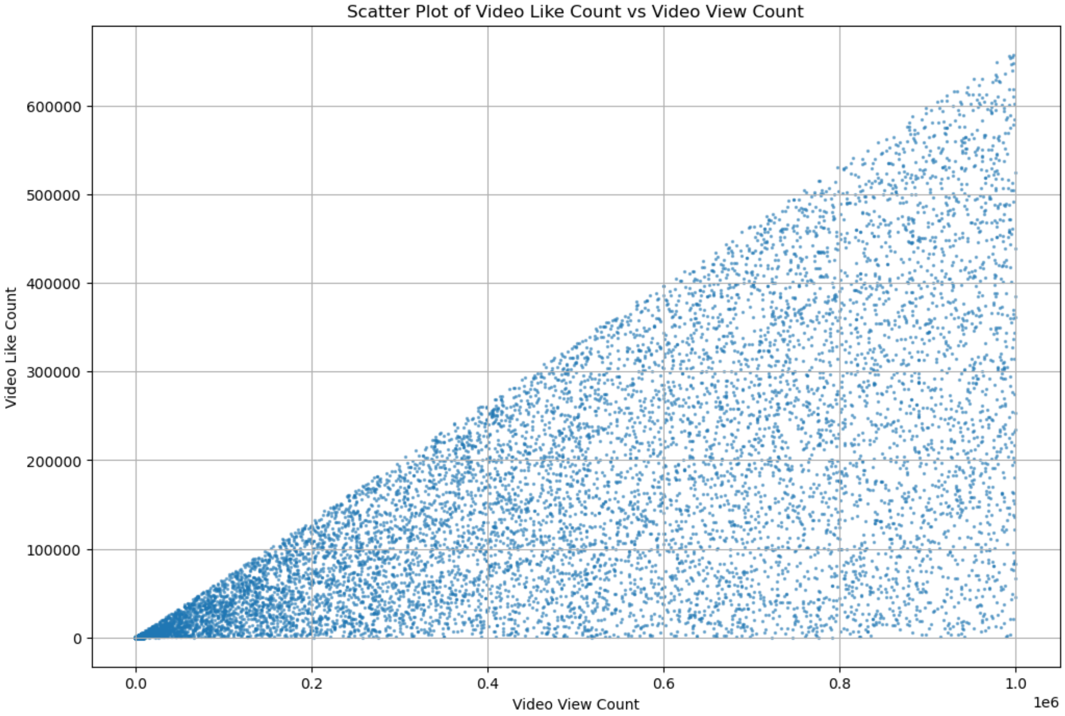


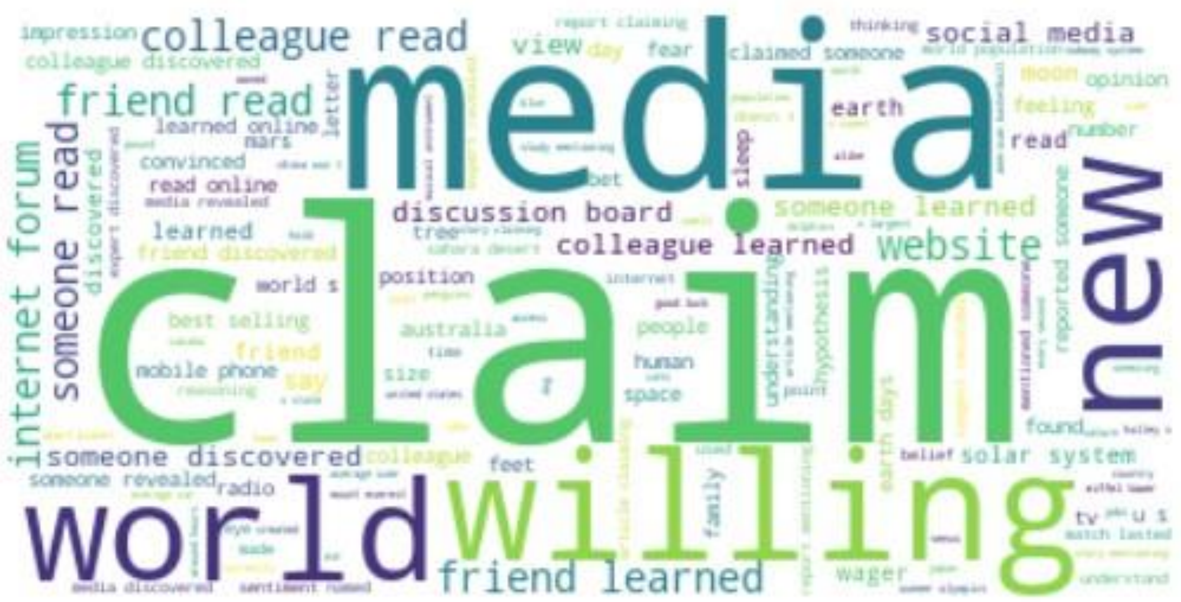


We selected the numeric columns in the DataFrame using the select_dtypes method and computed the correlation matrix of these numeric columns using the Corr method. Finally, we used seaborn’s heatmap function to plot the correlation matrix. The heatmap plot shows the correlation between all numeric variables in the dataset. The color in the heatmap shows how much two variables are correlated: a value close to 1 or -1 means a strong positive or negative correlation, respectively. A value close to 0 means no correlation.

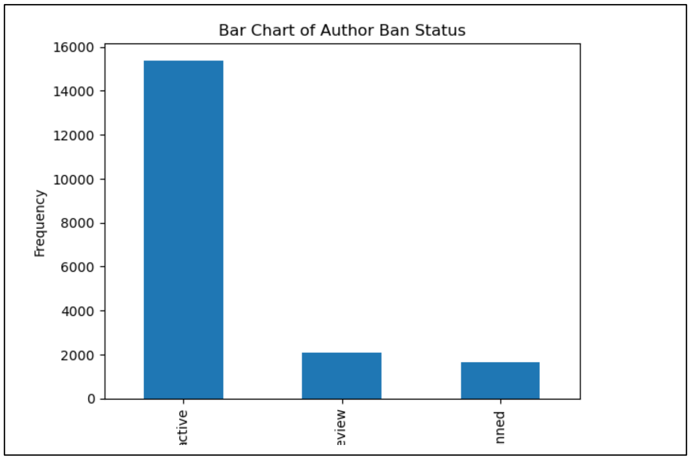

Bar Chart of Video Like Count by Claim Status
1e9
1.6
1.4
1.2
1.0
0.8
0.6
0.4
0.2
0.0
claim
opinion
claim_status

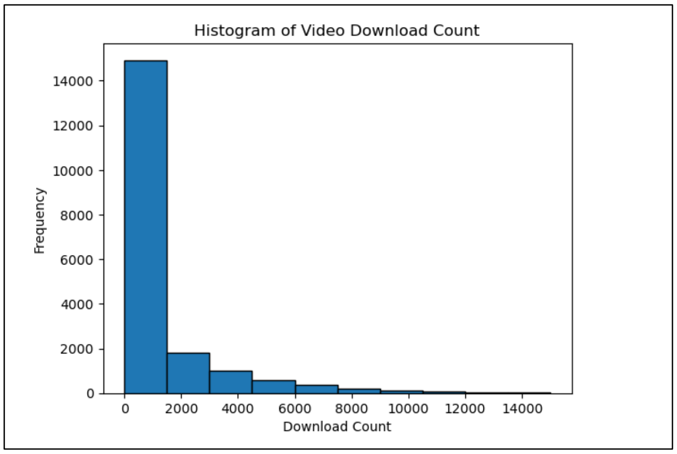

Pie Chart of Verified Status
not verified
count
93.7%
6.3%
verified

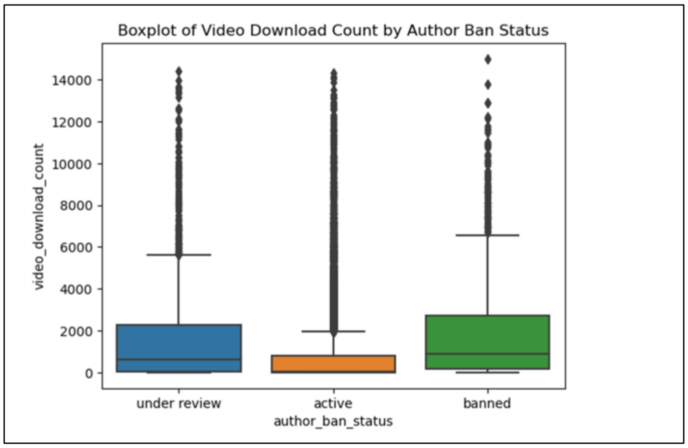

Scatter Plot of Video Like Count vs Video View Count
600000
500000
400000
300000
200000
100000
0
Video Like Count
0.0
0.2
0.4
0.6
0.8
1.0
1e6
Video View Count

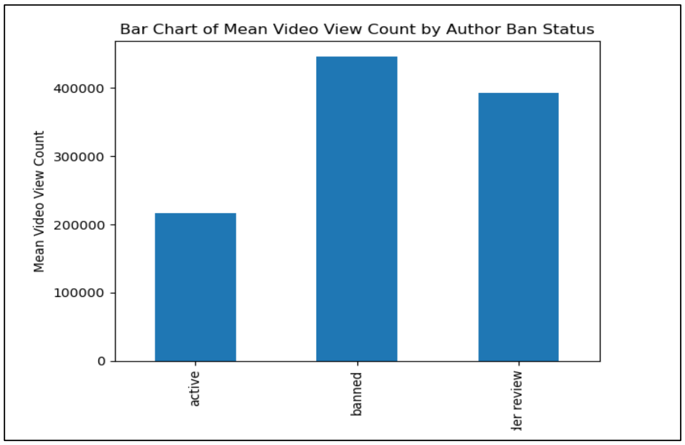


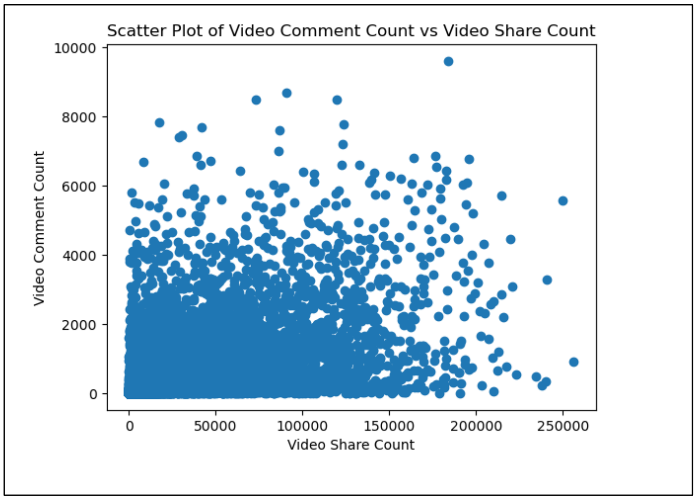
Word Cloud of Video Transcription Text

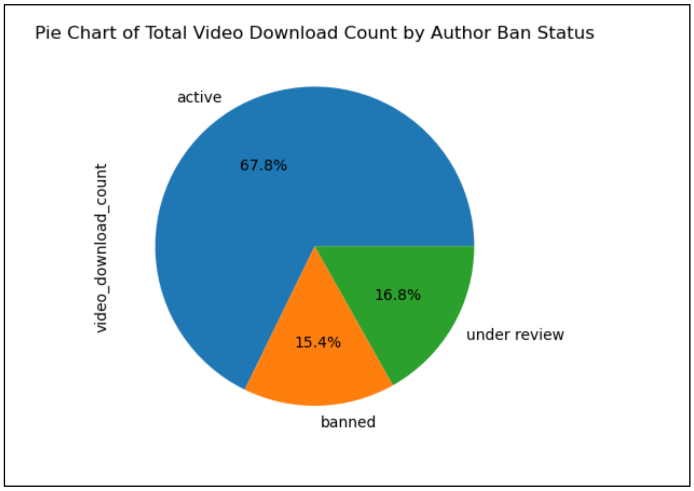


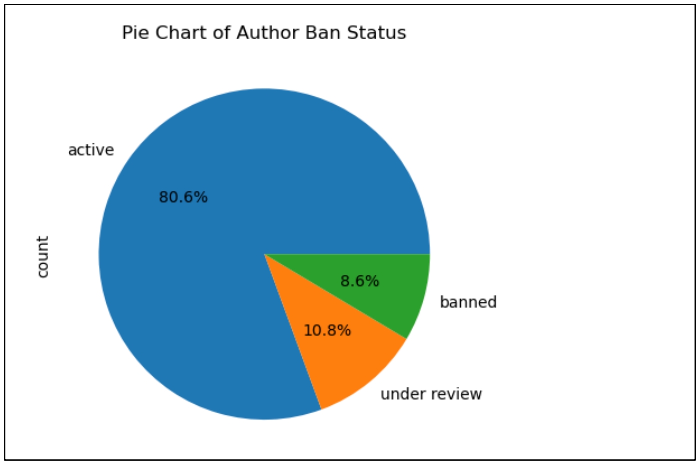








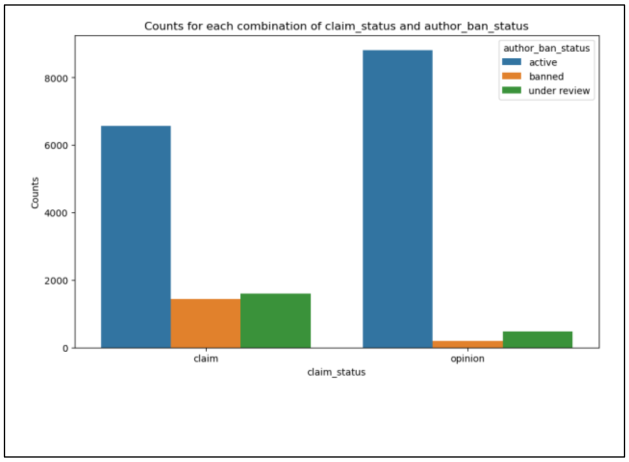


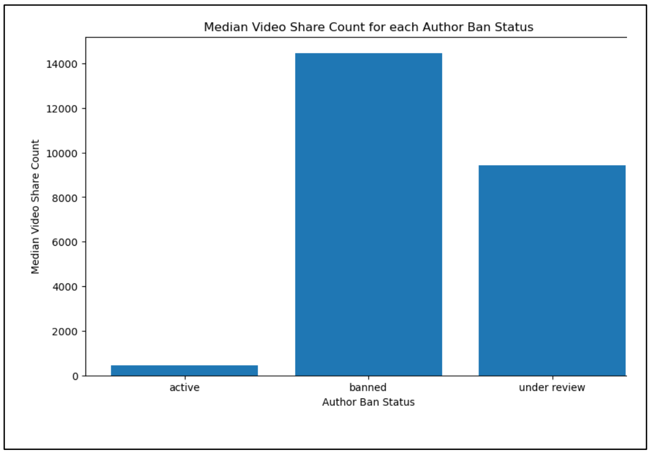


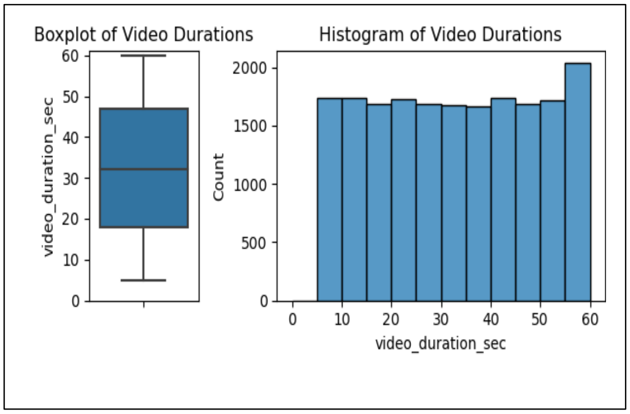
In this page we can see graphs of different variables visualizing the characteristics of the variables, such as its central tendency, spread, and shape of its distribution. We can also see some graphs showing the relationship between two variables. Such as the pie chart of total video download by author status.

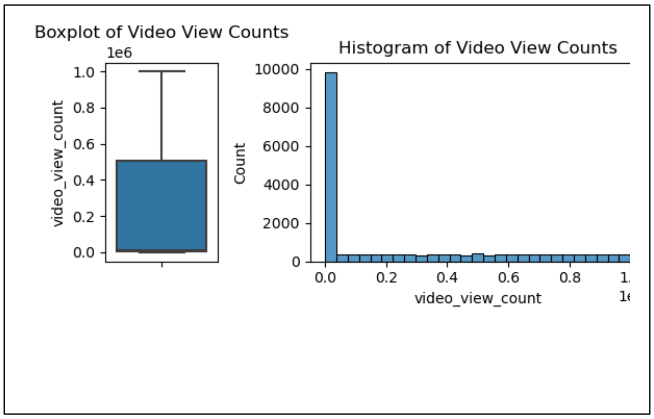
Data Engineering(25-30hours):

Data engineering includes data cleaning, preprocessing and Exploratory Data Analysis (EDA). For handling missing values we have used the df.isnull().sum() to check for missing values, Luckily I managed to get the tiktok_csv dataset with no missing values, which made this step relatively straightforward. So, we could jump into the next step which is EDA and calculations. We have performed a bunch of calculations on the dataset to understand the relationship between different variables better for further analysis. We have checked the distribution of claim status using value counts(). This gives an idea that "claim status" has two balanced classes. We have checked the counts of each group combination of **claim status** and **author ban status**. This helps understand the relationship between these two categorical variables. We have calculated the mean and median view counts for videos with a "claim" status and an "opinion" status. This could help identify trends or differences in engagement between these two categories. We have calculated the median share count grouped by **author ban status**. This gives insight into how ban status might affect user engagement. I have also plotted this calculation in various graphs to visualize the calculation and relationship better. Graphs are shown below.

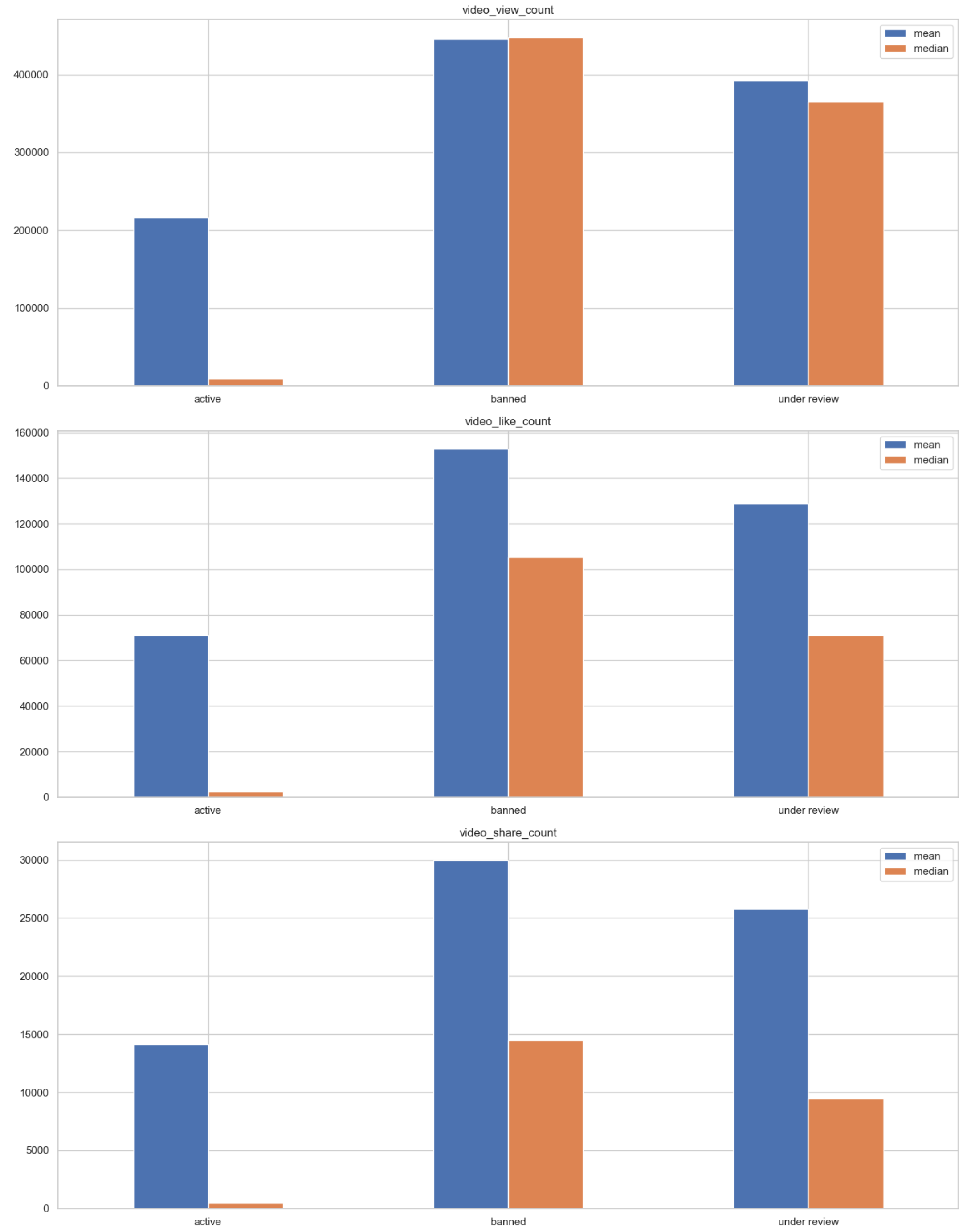






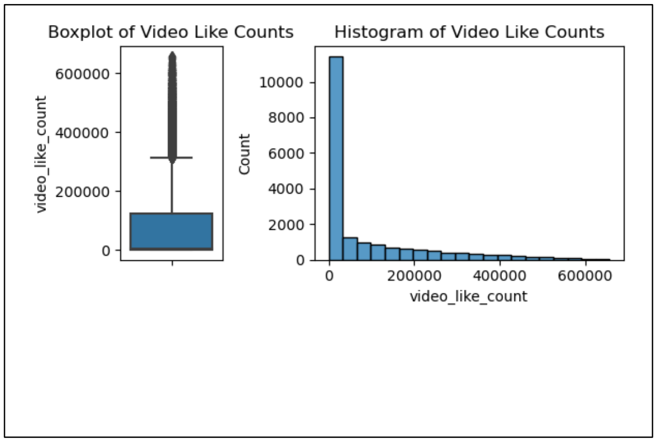


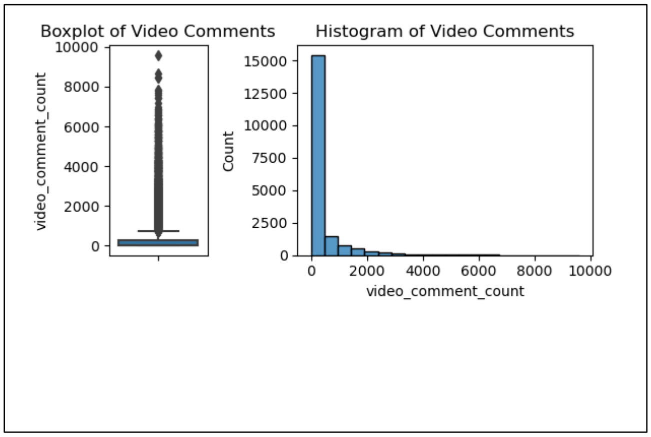
We have created new features “likes per view”, “comments per view”, “shares per view” using the feature engineering process to measure video engagement rates and then analyzing these rates across different claim statuses and author ban statuses.

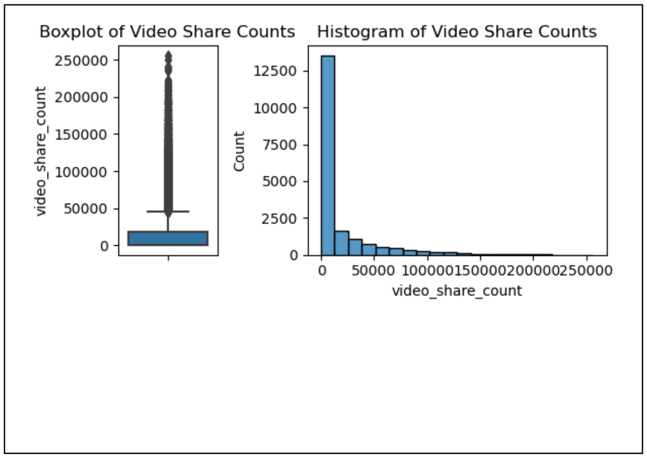


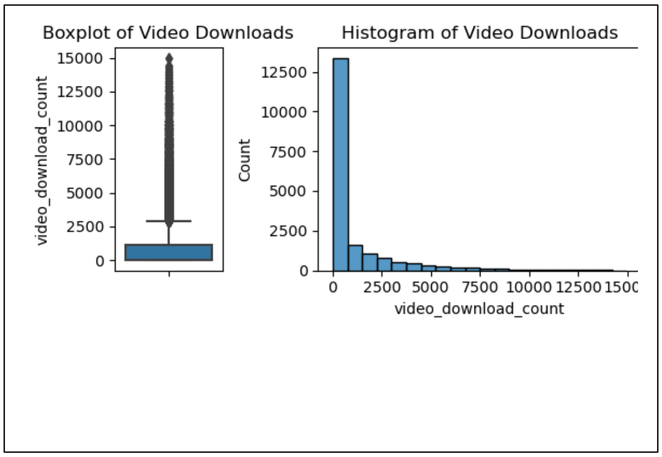
In this graph, we can see the mean and median values of video view count, video like count, and video share count vary with author ban status. The bar plots provide a visual representation of these relationships. It is apparent videos from banned authors or those under review tend to have higher engagement (views, likes, shares) compared to videos from active authors.

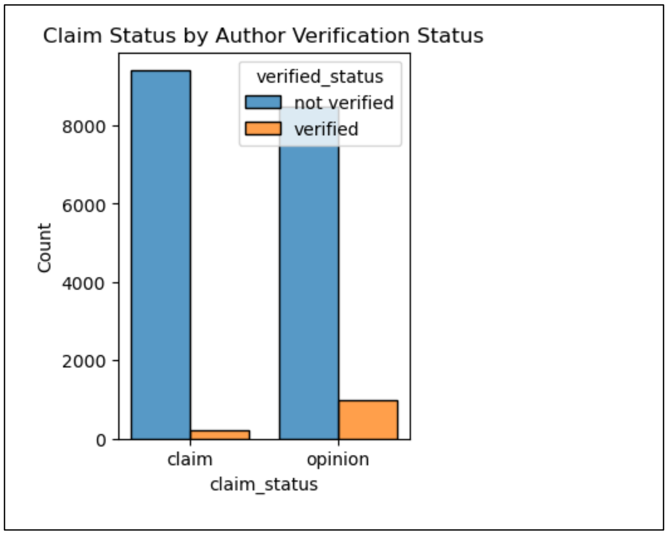
Below are some visualizations done to show the distribution of video durations, video like counts, video comments, video share and video downloads of users. The boxplots provide a summary of the distribution’s central tendency and spread, including any potential outliers, while the histograms provide a more detailed view of the distribution’s shape.

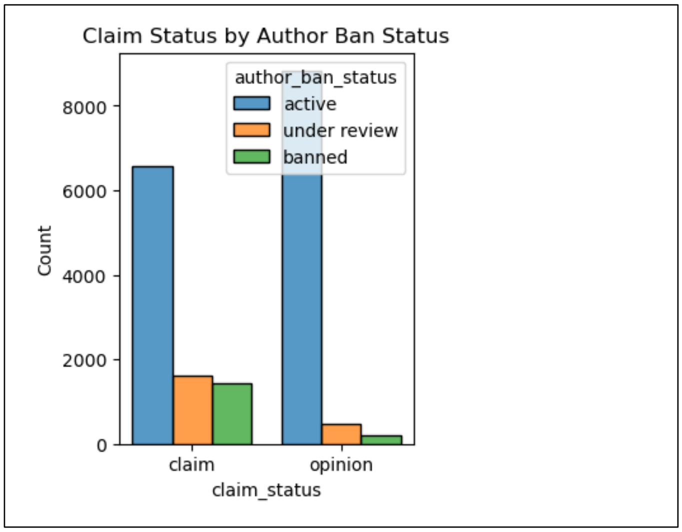


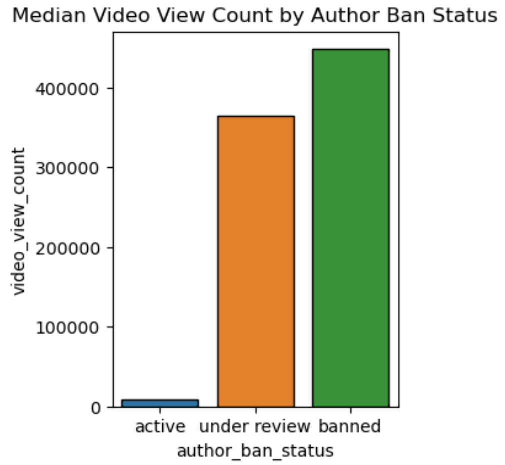


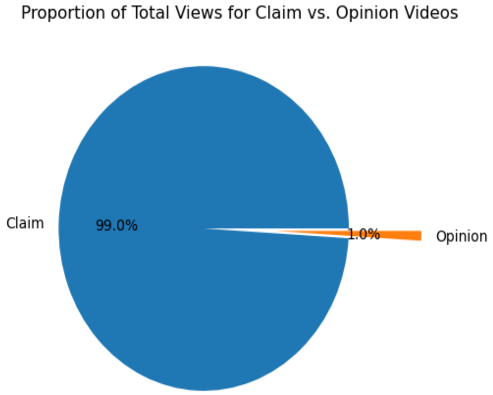


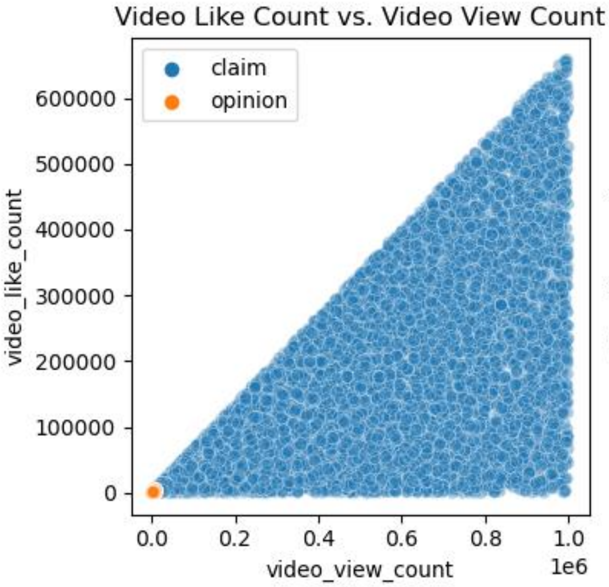


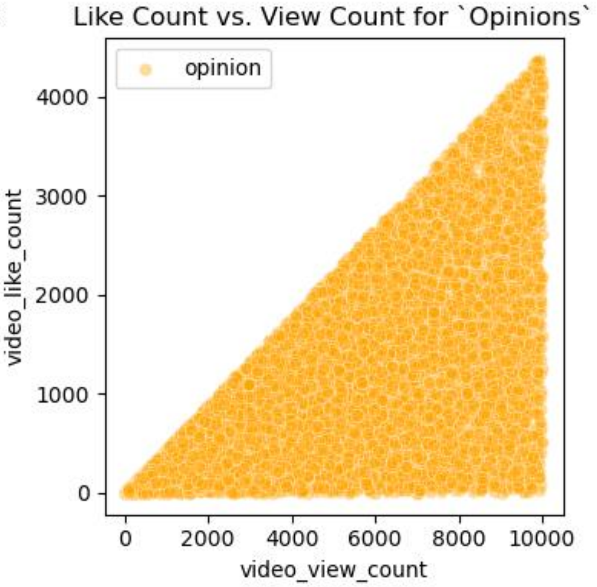






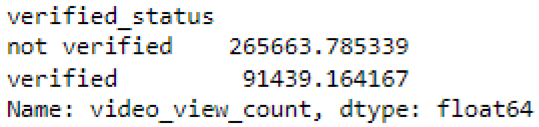






We have created a few more graphs visualizing the distribution of claim statuses with respect to ban status, claim status with respect to verification statues, video view counts with respect to ban status. These provide insights into the relationship between the mentioned variables. We have created a pie chart to show the Proportion of Total Views for Claim vs. Opinion Videos. We have also created scatter plots to show the relationship of video like counts vs video view counts according to claim status “claim” and “opinion.”

Statistical Analysis(10 hours): The goal of this part of the project is to find whether there’s a significant difference in the view counts of TikTok videos posted by verified accounts compared to those posted by unverified accounts.



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

Consequently, we propose the following null and alternative hypotheses:

Null Hypothesis (H0): There is no significant difference in the number of views for TikTok videos posted by verified accounts versus unverified accounts.

Alternative Hypothesis (Ha): There is a significant difference in the number of views for TikTok videos posted by verified accounts versus unverified accounts. We consider the variable video view count and divided them into two subgroups verified and not verified and calculate the mean of each subgroup. Then conducted the hypothesis testing to calculated the p-value which is 2.609e-120. This value is much smaller than the chosen significance level of 0.05 (or 5%). Therefore, the test strongly rejects the null hypothesis and concludes that there is a significant difference between the number of views for TikTok videos posted by verified accounts versus unverified accounts.

4. Deliverable (10 hours):

We can deliver a brief report to the audience in a PowerPoint presentation showing the key points of our analysis. The graphs can engage the audience and have them show interest rather than just looking at words, numbers and a bunch of codes. The entire project could be uploaded into the GitHub repository and the link to the repository could be sent to the client via email.