Phase 1: Project Proposal

TikTok Predictive Claim Model

OBJECTIVE

Design and construct a Machine Learning (ML) model with the ability to distinguish between videos containing factual claims and those

PLANNING and ANALYZING

	Milestone	Tasks	Outcome/Deliverables	Stakeholders	Timeline
PLAN	Milestone 1	• Establish structure for project workflow • State the Scope of the project • Define the final outcome needed • Gather data from different resources • Compile summary info on data	• Global Project Proposal • Data Ready to EDA	ALL	2 weeks
ANALYZE	Milestone 2	• Evaluate the model • Begin EDA • Data Exploration and Cleaning • Data Formatting	• EDA Report • Data Ready to Model	Data Science Team	4 weeks

CONSTRUCTING and EXECUTING

	Milestone	Tasks	Outcome/Deliverables	Stakeholders	Timeline
CONSTRUCT	Milestone 3	• Conduct hypothesis testing • Compute descriptive statistics • Build visuals • Build a regression model • Build machine learn model	• Tableau Dashboards • ML Final Model • Regression Model	Data Science Team	4 weeks
EXECUTE	Milestone 4	• Finalize results • Present findings w/stakeholders	• Executive Summary • Report	ALL	2 weeks

TikTok

Executive Summary

Phase 2: Understand the Data

PROBLEM

The TikTok data team seeks to develop a machine learning model to assist in the classification of claims for user submissions.

OBJECTIVE

The team performed a preliminary investigation of the claims classification dataset with the aim of learning important relationships between variables, trends and valuable insights.

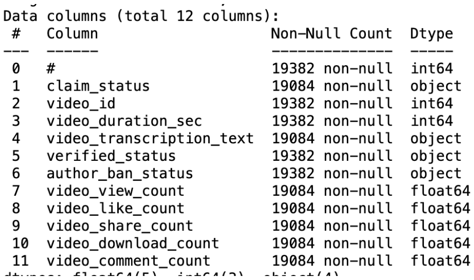
NEXT STEPS

The impact of this preliminary analysis will be evident in the next steps:

- Exploratory Data Analysis
- Statistical Tests
- Regression Modelling
- Machine Learning Models

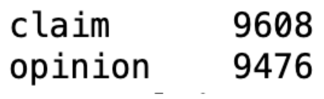
UNDERSTAND THE DATA

1) Out of the 19382 entries, some variables do have missing values, data cleaning will be needed for those columns.



```
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)
```

2) The counts of each claim status are quite balanced.

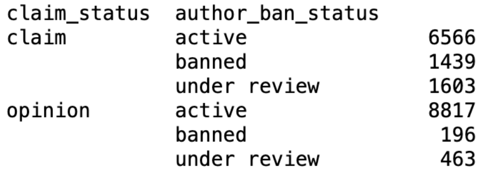


```
claim      9608
opinion    9476
```

3) The average view count of videos with "claim" status is 10x times higher than the average view count of videos with “opinion” status

- `Mean_view_count_claims: 501029.45`
- `Mean_view_count_opinions: 4956.43`

4) When the video status is marked as "claim" and the author is banned, there is a 7.5-fold increase in counts compared to instances where the author is banned but the video status is classified as "opinion.”



```
claim_status  author_ban_status
claim         active               6566
              banned               1439
              under review         1603
opinion       active               8817
              banned                196
              under review          463
```

KEY TAKEAWAYS

- **Balanced Distribution**: Opinions and claims in the dataset are almost equally represented.
- **Data Cleaning Requirement**: Certain variables will necessitate cleaning during the Exploratory Data Analysis (EDA) phase.
- **Impact on Views**: Videos labeled as "claim" attract 10 times more views compared to those tagged as "opinion.”
- **Outlier Indication**: The maximum values in some variables (columns 7,8,9,10,11) significantly exceed their respective 75th percentiles, suggesting the presence of potential outliers.

Executive Summary

Phase 3: Exploratory Data Analysis (EDA)

TikTok

PROBLEM

The TikTok data team seeks to develop a machine learning model to assist in the **classification of claims** for user submissions.

OBJECTIVE

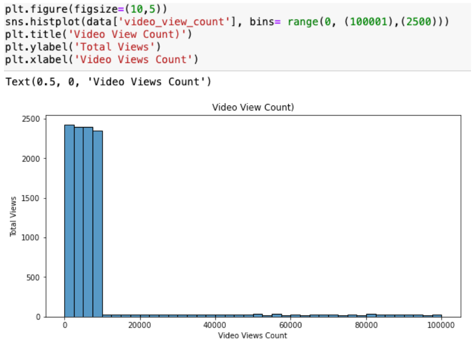
The task involves performing Exploratory Data Analysis (EDA) on a dataset using Python and Tableau, focusing on **data structuring, cleaning, outlier detection** and **visualization**. The analysis features graphs and boxplots analyzing key metrics such as *video duration, likes, comments, views, claim/opinion counts, and author ban statuses.*

NEXT STEPS

The EDA provided insights into the data's features, removed outliers, and revealed overall trends. Upcoming project phases will focus on:

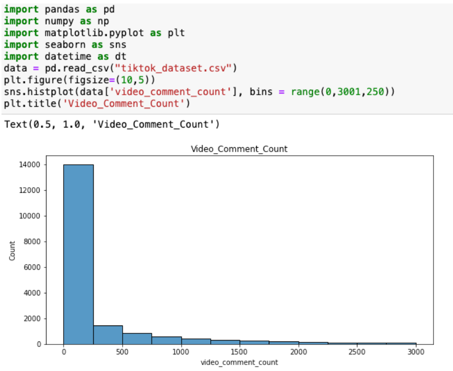
- Statistical Tests
- Regression Modelling
- Machine Learning Models

UNDERSTAND THE DATA



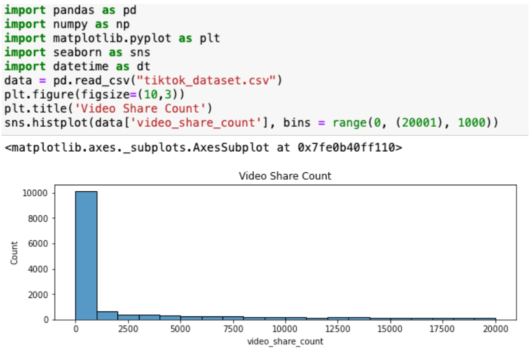
```
plt.figure(figsize=(10,5))
sns.histplot(data['video_view_count'], bins= range(0, (100001),(2500)))
plt.title('Video View Count)')
plt.ylabel('Total Views')
plt.xlabel('Video Views Count')
```

```
Text(0.5, 0, 'Video Views Count')
```



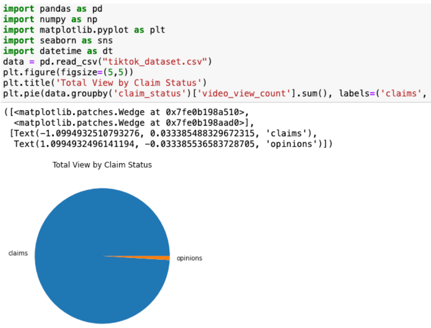
```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import datetime as dt
data = pd.read_csv("tiktok_dataset.csv")
plt.figure(figsize=(10,5))
sns.histplot(data['video_comment_count'], bins = range(0,3001,250))
plt.title('Video_Comment_Count')
```

```
Text(0.5, 1.0, 'Video_Comment_Count')
```



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import datetime as dt
data = pd.read_csv("tiktok_dataset.csv")
plt.figure(figsize=(10,3))
plt.title('Video Share Count')
sns.histplot(data['video_share_count'], bins = range(0, (20001), 1000))
```

```
<matplotlib.axes._subplots.AxesSubplot at 0x7fe0b40ff110>
```



```
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import seaborn as sns
import datetime as dt
data = pd.read_csv("tiktok_dataset.csv")
plt.figure(figsize=(5,5))
plt.title('Total View by Claim Status')
plt.pie(data.groupby('claim_status')['video_view_count'].sum(), labels=('claims',
```

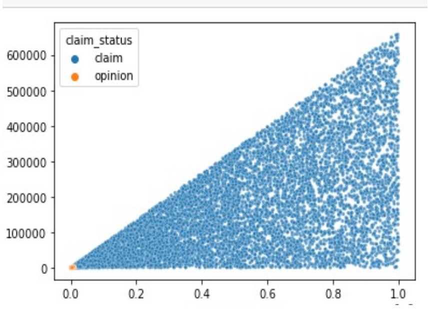
```
([<matplotlib.patches.Wedge at 0x7fe0b198a510>,
  <matplotlib.patches.Wedge at 0x7fe0b198aad0>],
 [Text(-1.0994932510793276, 0.033385488329672315, 'claims'),
  Text(1.0994932496141194, -0.033385536583728705, 'opinions')])
```

The data distribution is notably **right-skewed**, with a concentration the majority of data points in the lower 25% percentile, as observed variables like `video_view_count`, `video_share_count`, and `video_comment_count`.

Additionally, the 'Total View by Claim Status' pie chart reveals a significant insight: *a vast majority of views are associated with vide that have claims as their comment status.*

KEY INSIGHTS

- Viral videos, with high engagement, are usually 10 to 60 sec long.
- Videos with claims in comments garner most views.
- Over half of the videos get less than 100,000 views, indicating a skewed view count distribution.
- Presence of over 200 nulls across seven columns suggests incorporating these in future models for accurate insights.



claim_status
claim
opinion