PACE Strategy Document

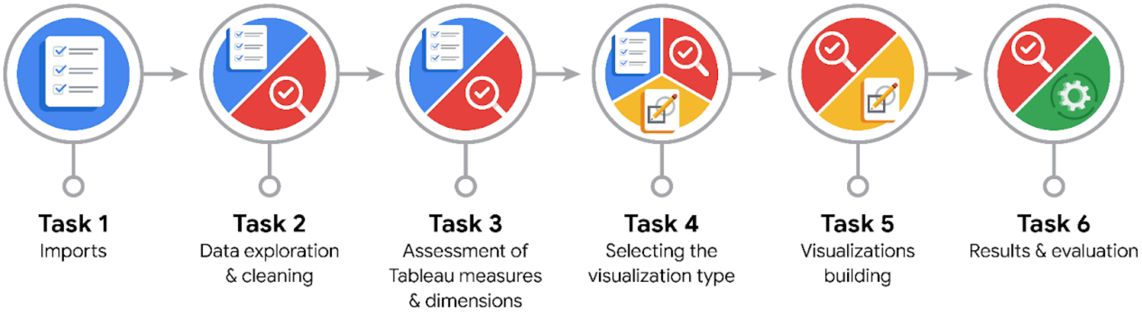
Phase 3: Advanced Data Visualization & Outlier Analysis for TikTok Claims Classification

Instructions

I will use this **PACE strategy document** to record my decisions and reflections as I work through this Phase-3 project. This document will serve as a guide, helping me consider my responses and reflections at different stages of the data analytical process.

Reference Guide

This project has six tasks; the visual below identifies how the stages of PACE are incorporated across those tasks.


Task 1
Imports
Task 2
Data exploration & cleaning
Task 3
Assessment of Tableau measures & dimensions
Task 4
Selecting the visualization type
Task 5
Visualizations building
Task 6
Results & evaluation

Data Project Questions & Considerations

PACE: Plan Stage

- What are the data columns and variables and which ones are most relevant to your deliverable?

There are 12 data columns and 19,383 rows in this dataset.

# (int) -TikTok assigned number for video with claim/opinion.

claim_status (obj): Specifies if the video contains an "opinion" or a "claim."
video_id (int): Unique identifier for each TikTok video.
video_duration_sec (int): Length of the video in seconds.
video_transcription_text (obj): Text transcription of the spoken content in the video.
verified_status (obj): Indicates if the user is verified ("verified" or "not verified").
author_ban_status (obj): Status of the user as “active,” “under scrutiny,” or “banned.”
video_view_count (float): Total number of views the video received.
video_like_count (float): Total number of likes the video received.
video_share_count (float): Total number of shares the video received.
video_download_count (float): Total number of downloads the video received.
video_comment_count (float): Total number of comments on the video.

The ones most relevant to me are the claim_status, author_ban_status, the engagement metric fields such as video_like_count, video_view_count, video_comment_count, etc fields. Additionally verified_status

- What units are your variables in?

video_duration_sec: Measured in seconds.

Most fields represent counts, except for **video_duration_sec**, which is explicitly measured in time units (seconds).

The other fields fall into two broad categories:

Identifiers:

#: TikTok assigned number for video with claim/opinion.

video_id: A unique identifier assigned to each TikTok video.

Textual Information or Attributes:

claim_status: Describes whether the video contains an "opinion" or a "claim."

video_transcription_text: Contains the transcribed text of the spoken content in the video.

verified_status: Indicates whether the author is "verified" or "not verified."

author_ban_status: Specifies the status of the user as "active," "under scrutiny," or "banned."

- What are your initial presumptions about the data that can inform your EDA, knowing you will need to confirm or deny with your future findings?

My initial assumption was that videos with a 'claim_status' classified as claims would have higher engagement metrics than those tagged as opinions, and so far, this is proving to be correct based on the correlations I have observed.

- Is there any missing or incomplete data?

Yes, the videos with missing values in the **claim_status** field are the same ones that also have missing values in the following fields:

Video_view_count

Video_like_count

Video_share_count

Video_download_count

Video_comment_count

video_transcription_text

- Are all pieces of this dataset in the same format?

some string (object) fields for categorical or textual information. Here's the breakdown:

Numerical Data:

int (Integer type):

#: TikTok-assigned number for videos with claims/opinions.

video_id: Random identifier assigned to each TikTok video.

video_duration_sec: Length of the video in seconds.

float (Floating-point type):

video_view_count: Total number of views the video received.

video_like_count: Total number of likes the video received.

video_share_count: Total number of shares the video received.

video_download_count: Total number of downloads the video received.

video_comment_count: Total number of comments on the video.

String Data (obj type):

claim_status: Indicates whether the video contains an "opinion" or a "claim."

video_transcription_text: Transcribed text of the words spoken in the video.

verified_status: Specifies if the user is "verified" or "not verified."

author_ban_status: Indicates the user’s permissions status: "active," "under scrutiny," or "banned."

This structure highlights that while the dataset is primarily numerical, the inclusion of categorical and textual data enriches its potential for detailed analysis

- Which EDA practices will be required to begin this project?

The EDA practice of Discovering and structuring.

PACE: Analyze Stage

- What steps need to be taken to perform EDA in the most effective way to achieve the project goal?

To perform EDA effectively and achieve the project goal, the following steps were taken:

1. Discovering: Initial exploration was conducted to understand the dataset, identify patterns, trends, and anomalies, and get a sense of the data's overall structure.

2. Structuring: The dataset was organized to enhance its usability by grouping relevant observations for better analysis and deeper insights.

3. Validating: This step was iteratively performed throughout the process to ensure data accuracy, reliability, and consistency, addressing issues such as missing values, outliers, or inconsistencies whenever they were identified.

4. Presenting: Visualizations were created using both Python and Tableau to analyze and communicate the structure of the data effectively. These visualizations highlighted key patterns, trends, and insights, aiding in better decision-making and aligning with the project's objectives.

This systematic approach ensured that the dataset was thoroughly explored, well-structured, and actionable insights were drawn.

- Do you need to add more data using the EDA practice of joining? What type of structuring needs to be done to this dataset, such as filtering, sorting, etc.?

The dataset is fairly representative of the cause being addressed, but augmenting it with additional data could provide deeper insights. For instance, adding demographic information about the users could enhance the analysis.

As for structuring, the dataset is in good shape overall, but some improvements could be made:

Filtering: Rows with missing values could be filtered out or imputed, depending on the context and importance of the missing data.

Sorting: The data can be sorted based on various engagement metrics, such as video views, likes, shares, or comments, to prioritize analysis of the most or least engaging content.

Grouping: The dataset can be grouped by specific attributes, such as claim status, verified status, or author ban status, to identify patterns, trends, or differences within these groups.

These steps would further refine the dataset, making it more structured, insightful, and actionable for analysis.

- What initial assumptions do you have about the types of visualizations that might best be suited for the intended audience?

Bar charts and pie charts are well-suited for the intended audience, as they effectively communicate categorical data and proportions. Additionally, visualizations like histograms, scatterplots, and heatmaps can provide a deeper understanding of the data, offering insights into distribution patterns, relationships, and correlations within the dataset.

PACE: Construct Stage

- What data visualizations, machine learning algorithms, or other data outputs will need to be built in order to complete the project goals?

To complete the project goals, data visualizations such as scatter plots and box plots will be valuable for identifying relationships and detecting outliers. Additionally, machine learning algorithms like logistic regression and Support Vector Machines (SVM) can be used to classify and predict outcomes based on the dataset. Other important data outputs include summary statistics (mean, median, standard deviation), correlation matrices to identify relationships between variables, and trend analysis to highlight significant patterns over time. These outputs, combined with visualizations and machine learning results, will help in drawing meaningful insights and making informed decisions based on the data.

- What processes need to be performed in order to build the necessary data visualizations?

To build the necessary data visualizations, the following processes should be performed: Data Structuring: As a preprocessing step, the data needs to be organized, cleaned, and transformed to ensure it's in a suitable format for visualization. This may involve handling missing values, filtering, and aggregating data as needed. Data Validation: Before creating visualizations, the data should be validated to ensure its accuracy and consistency. This includes verifying data types, checking for errors, and confirming that all relevant variables are included. Data Formatting: In both Python and Tableau, the data needs to be properly formatted, ensuring it aligns with the requirements of the specific visualization tools. This includes converting categorical variables to appropriate formats, creating necessary calculated fields, and ensuring the data is structured for the intended analysis. Once these steps are complete, the visualizations can be created using Python libraries (such as Matplotlib or Seaborn) and Tableau for interactive exploration of the data.

- Which variables are most applicable for the visualizations in this data project?

The variables most applicable for the visualizations in this data project include **claim_status**, **author_ban_status**, and the **video engagement metrics** (such as video_view_count, video_like_count, video_share_count, video_download_count, and video_comment_count). These variables will help illustrate patterns and trends related to content claims, user behavior, and engagement, making them key to understanding the dynamics of the videos in the dataset.

- Going back to the Plan stage, how do you plan to deal with the missing data (if any)?

To deal with the missing data, specifically the observations where claim_status is missing, I would consider the following options:

Inform the Concerned Stakeholders:

I would notify the relevant team or data provider about the missing data for claim_status to understand the reason behind it and determine if it's a data collection issue that can be addressed. This would help in understanding if there's a systematic issue causing the missing data and if there's a possibility to gather the missing information.

Create a New Category:

If the missing claim_status can be reasonably attributed to uncertainty or incomplete information, I could categorize those videos as "Unclassified" or another appropriate label. This would allow for the inclusion of the data while acknowledging the uncertainty and maintaining the integrity of the analysis.

Remove the Rows:

If the missing data constitutes a significant portion of the record (with other variables missing as well), I may decide to exclude these rows from analysis. This step would ensure the quality and consistency of the dataset, especially if the missing values are pervasive and removing them does not heavily impact the analysis.

Impute Missing Values for Numerical Data:

For the numerical data (such as video_view_count, video_like_count, etc.), I would consider imputing the missing values using the mean, median, or forward/backfilling, depending on the data's distribution and the nature of the missingness.

Handle Missing Textual Data:

For textual columns like video_transcription_text, I could either:

Fill the missing text with a placeholder (e.g., "No transcription available").

Use NLP techniques for text prediction (if feasible).

Remove rows with missing text, particularly if it's essential to the analysis.

The choice of method would depend on the context, the amount of missing data, and the impact of including or excluding these records. By considering these options, I aim to balance the dataset's completeness and the quality of the analysis while ensuring accurate and meaningful insights.

PACE: Execute Stage

- What key insights emerged from your EDA and visualizations(s)?

The key insight that emerged from the EDA and visualizations is that claim_status is strongly correlated with video_view_count and other engagement metrics, such as video_like_count, video_share_count, and video_comment_count. Videos categorized as claims tend to have significantly higher engagement, suggesting that claim videos, often more controversial or attention-grabbing, generate more reactions from users.

- What business and/or organizational recommendations do you propose based on the visualization(s) built?

Based on the visualizations, I recommend the following:

Prioritize Engagement Metrics: Focus on metrics like views, likes, shares, and comments when building a classification model to optimize content for higher engagement.

Consider User Classifications: Incorporate user status (e.g., banned, active, verified) to understand how user classification impacts video success.

Segment Data for Targeted Strategies: Segment data by engagement and user status to tailor content distribution and improve interaction.

Content Optimization: Analyze trends to provide creators with recommendations on video duration, content type, and engagement practices to boost reach and interaction.

- Given what you know about the data and the visualizations you were using, what other questions could you research for the team?

I could research the correlations between various video engagement metrics, such as likes, shares, and comments, to identify which factors most strongly influence engagement. Additionally, I could investigate how video duration relates to engagement metrics, as longer or shorter videos may impact viewer interaction differently. This analysis could help optimize video content strategies.

- How might you share these visualizations with different audiences?

I would share viewer-friendly visualizations, such as bar charts and pie charts, with stakeholders or those without extensive data analysis experience. For more complex visualizations, such as scatter plots or heatmaps, I would share them with data professionals or cross-functional teams who are comfortable interpreting such data. These visualizations can be shared through various formats, including PowerPoint presentations, dashboards, and detailed reports.