

TikTok PACE Strategy Document III

Exploratory Data Analysis

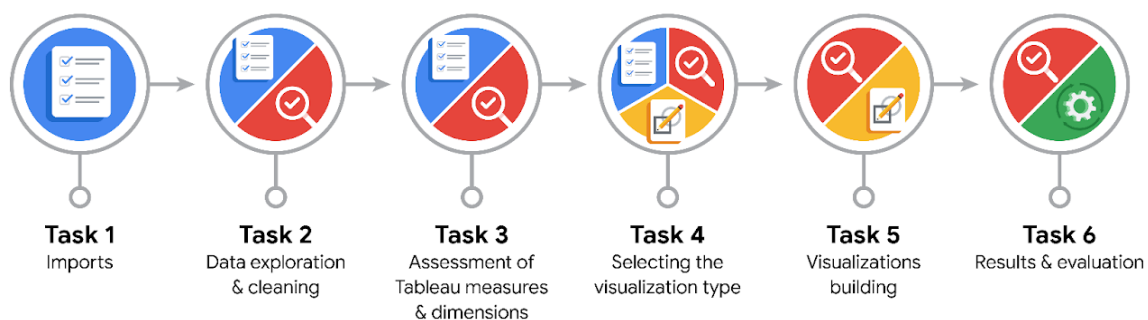
Introduction

PACE stands for Plan, Analyze, Construct and Execute. It is a framework that illustrates the foundation and structure for data analysis projects and each letter represents an actionable stage in a project. The stage “Plan” involves the definition of the project scope, the research of business data and the workflow development. The stage “Analyze” involves data scrubbing, data conversion and database formatting. The stage “Construct” involves building models and machine learning algorithms and selecting a modeling approach. The stage “Execute” involves the presentation of results to decision-makers, stakeholders and others in order to receive feedback. This framework is built upon an iterative cycle where each stage may reveal new insights, requiring the return to earlier stages. A PACE strategy document is used to record decisions and reflections at different stages of the data analytical process. It typically includes the definitions of roles and actions to ensure clarity and accountability.



Purpose

TikTok users have the ability to report videos and comments that contain user claims. These reports identify content that needs to be reviewed by moderators. This process generates a large number of user reports that are difficult to address quickly. TikTok is working on the development of a predictive model that can determine whether a video contains a claim or offers an opinion. With a successful prediction model, TikTok can reduce the backlog of user reports and prioritize them more efficiently. For this stage of the project, we identify six main tasks that are presented in the following visual.





Considerations



PACE: Plan Stage

- What methods are best for identifying outliers?

We use *NumPy* to calculate the mean and median and understand the data range. We create boxplots and histograms to visualize and observe how the data is distributed. We also create scatter plots which allow us to observe trends, patterns and outliers between variables.

- How do we decide whether to keep or remove outliers from future models?

There are three main ways to handle outliers: retaining them, removing them or reassigning values. The choice depends on the nature of the outliers and the model's assumptions. If the outliers are clearly errors like typos and the data will be used in modeling or machine learning, removing them is often best (though this is the least common option). If the dataset is small or used for modeling, reassigning values might be necessary. If the data will be used for EDA or the model is resistant to outliers, it is usually best to keep them.

- Which columns and variables in the dataset are most critical for our analysis and final deliverable?

Relevant variables include the video duration in seconds, the number of views, the number of likes, the number of comments, the number of shares, the number of downloads, the claim status (claim or opinion), the author ban status (active, under review, banned) and the user verification status (verified, unverified). These variables are vital for understanding user engagement, content performance and the distinction between claims and opinions.

- What are the units of measurement for each variable?

Video duration is measured in seconds, the different counts (views, likes, comments, shares, downloads) are recorded based on user interaction and the different statuses (claim, author ban, verification) contain categorical values.

- What are our initial assumptions about the data that will guide our exploratory data analysis and what do we expect to confirm or adjust through our findings?

Engagement metrics (views, likes, comments) will likely have right-skewed distributions, meaning most videos get low engagement, with a few highly popular outliers. Claim videos will likely have higher view counts and engagement rates compared to opinion videos. Non-active authors (banned or under review) might post content that attracts more views, especially in the claim category.

- Is there any missing or incomplete data in the dataset?

There are over 200 null values across 7 different columns. It is critical to investigate the cause of the missing data and decide whether to fill, exclude or impute these values in the analysis.

- Is the dataset consistent in terms of format?

The dataset seems generally consistent in terms of format, but the presence of null values could cause some inconsistencies. It is important to ensure all variables are in the correct data type and that there are no mismatches (e.g. strings instead of numeric values).

- What exploratory data analysis techniques will we need to use to start this activity?

Null values should be handled and consistent data formatting should be ensured. Means, medians and standard deviations for the variables should be calculated. Boxplots and histograms should be created to assess data distribution and outliers. Relationships between the claim status and the engagement variables (views, likes, comments) should be examined.




PACE: Analyze Stage

- What actions should be taken during EDA to ensure the project reaches its objectives?

We should investigate the distribution of variables using histograms and boxplots, check for and handle outliers and missing values, analyze the relationship between the claim status and the engagement metrics (views, likes, shares) and conduct correlation analysis to identify key predictors for the claim classification model.

- Is there a need to combine more datasets as part of the EDA process and what structuring tasks (such as filtering or sorting) are necessary for the current data?

If additional data on user demographics, video content categories or geographic location becomes available, combining these datasets could enhance the analysis. For the current dataset, structuring tasks like filtering by the claim status and sorting by the view count or the author ban status will help organize the analysis.

- 
- What initial ideas do we have about which types of visualizations might be most effective for our audience?

Bar charts will be ideal for comparing counts of claims and opinions. Boxplots can be created to show the distribution of key engagement metrics and identify outliers. Stacked bar charts can visualize the breakdown of engagement metrics by the claim status and the author ban status. Histograms will be valuable for examining the distribution of view and like counts.



PACE: Construct Stage

- What visualizations, machine learning models or other outputs will be essential to achieve the project's goals?

Box plots, histograms, scatter plots, bar charts and possibly a classification model to predict the claim status.

- What steps are needed to create the necessary visualizations?

Cleaning and structuring the dataset by addressing missing values, filtering outliers and generating visualizations like boxplots, histograms and stacked bar charts.

- Which variables should be prioritized in the data visualizations for this project?

Engagement variables (views, likes, shares etc.) and the author ban status should be prioritized in the data visualizations to examine their relationship with the claim status.

- Returning to the planning stage, how will we address any missing data that is identified?

We should analyze the source of missing data. If the missing data is random, we may fill it with averages/medians. However, if it is systematic, we can consider excluding or flagging it in the analysis.



PACE: Execute Stage

- What key insights have we gathered from your EDA and visualizations?

Most videos have low engagement, with a few popular outliers. This suggests the need for log transformations or non-parametric models. Claim videos consistently receive more views than opinion videos, confirming the assumption that claims tend to attract more attention. Non-active authors (banned or under review) tend to have higher engagement, particularly with claim videos.

- What recommendations would we provide to the organization based on our findings?

By focusing on claim videos, we can better understand content virality on TikTok. The behaviors of non-active authors and their impact on user engagement should be investigated to shape content moderation policies. Engagement metrics should be considered as primary indicators for identifying claim content in the classification model.

- Based on our current knowledge of the data and visualizations, what additional research questions could be investigated for the team?

Key research questions include:

- What characteristics lead to a video being flagged as a claim versus an opinion?
- How does the engagement differ by video content genre or length, beyond the current focus on claim vs. opinion?
- Can demographic data of authors/viewers further explain engagement trends?

- How do we plan to present these visualizations to different stakeholders or audiences?

Dashboards like Tableau provide a clear summary of key trends and patterns, tailoring the presentation for managers or technical audiences. In general, for non-technical audiences, we should use simplified bar charts and accessible Tableau dashboards with clear explanations. For data-savvy stakeholders, we can present detailed boxplots, histograms and correlation analyses to explain deeper insights into data patterns.