**Course Three**

# Go Beyond the Numbers: Translate Data into Insights



# Instructions

Use this PACE strategy document to record decisions and reflections as you work through this end-of-course project. You can use this document as a guide to consider your responses and reflections at different stages of the data analytical process. Additionally, the PACE strategy documents can be used as a resource when working on future projects.

# Course Project Recap

Regardless of which track you have chosen to complete, your goals for this project are:

* Complete the questions in the Course 3 PACE strategy document
* Answer the questions in the Jupyter notebook project file
* Clean your data, perform exploratory data analysis (EDA)
* Create data visualizations
* Create an executive summary to share your results

# Relevant Interview Questions

Completing the end-of-course project will help you respond to these types of questions that are often asked during the interview process:

* How would you explain the difference between qualitative and quantitative data sources?
* Describe the difference between structured and unstructured data.
* Why is it important to do exploratory data analysis?
* How would you perform EDA on a given dataset?
* How do you create or alter a visualization based on different audiences?
* How do you avoid bias and ensure accessibility in a data visualization?
* How does data visualization inform your EDA?

**Reference Guide**

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



**Data Project Questions & Considerations**

**PACE: Plan Stage**

* What are the data columns and variables and which ones are most relevant to your deliverable?
*  **claim\_status**: Important for identifying whether the video is an opinion or a claim.
*  **video\_transcription\_text**: Useful for analyzing the content of the videos.
*  **verified\_status**: Indicates the credibility of the source.
*  **author\_ban\_status**: Provides context on the author's standing within the TikTok community.
*  **video\_view\_count**, **video\_like\_count**, **video\_share\_count**, **video\_download\_count**, **video\_comment\_count**: Metrics that indicate the engagement and reach of the videos.
* What units are your variables in?
* # int64
* claim\_status object
* video\_id int64
* video\_duration\_sec int64
* video\_transcription\_text object
* verified\_status object
* author\_ban\_status object
* video\_view\_count float64
* video\_like\_count float64
* video\_share\_count float64
* video\_download\_count float64
* video\_comment\_count float64
* dtype: object
* 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?
*  **Claim Status Distribution**: There may be an unequal distribution between "opinions" and "claims."
*  **Video Engagement**: Verified users' videos might have higher engagement (views, likes, shares, comments) compared to non-verified users.
*  **Author Ban Status**: Banned authors might have videos with distinct engagement patterns compared to active authors.
*  **Video Duration**: Most videos will likely fall within a typical TikTok duration range (e.g., 15-60 seconds).
* Is there any missing or incomplete data?
* # 0
* claim\_status 298
* video\_id 0
* video\_duration\_sec 0
* video\_transcription\_text 298
* verified\_status 0
* author\_ban\_status 0
* video\_view\_count 298
* video\_like\_count 298
* video\_share\_count 298
* video\_download\_count 298
* video\_comment\_count 298
* dtype: int64

Yes, video\_id, video\_duration\_sec, verified\_status and author\_ban\_status are the only data types are the ones that are non-null values.

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

No, the dataset contains columns with different data types:

* **Integer**: #, video\_id, video\_duration\_sec
* **Object (String)**: claim\_status, video\_transcription\_text, verified\_status, author\_ban\_status
* **Float**: video\_view\_count, video\_like\_count, video\_share\_count, video\_download\_count, video\_comment\_count
* Which EDA practices will be required to begin this project?
*  **Clean the Data**: Address missing values and ensure correct data types.
*  **Explore Data Distributions**: Analyze distributions of numerical variables and categorical counts.
*  **Examine Relationships**: Investigate correlations and interactions between variables.
*  **Text Analysis**: Review the video\_transcription\_text for insights or patterns.

**PACE: Analyze Stage**

* What steps need to be taken to perform EDA in the most effective way to achieve the project goal?
*  **Data Cleaning**: Handle missing values, correct data types, and remove duplicates.
*  **Descriptive Analysis**: Compute summary statistics and visualize distributions for numerical and categorical variables.
*  **Correlation Analysis**: Identify relationships between numerical variables and check for correlations.
*  **Categorical Analysis**: Analyze and visualize the distribution of categorical variables like claim\_status, verified\_status, and author\_ban\_status.
*  **Text Analysis**: Perform basic text analysis on video\_transcription\_text to identify common themes or patterns.
*  **Engagement Metrics**: Compare engagement metrics (views, likes, shares, comments) across different categories and statuses.
* 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.?

To analyze engagement metrics, compare the average views, likes, shares, and comments for videos categorized by claim\_status, verified\_status, and author\_ban\_status. This comparison will help identify patterns and differences in engagement based on these factors, revealing how they might influence video popularity and interaction.

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

For the intended audience, visualizations such as bar charts for categorical comparisons, histograms for distribution analysis, and scatter plots for relationships between engagement metrics would be effective. Heatmaps could also be useful for examining correlations, while word clouds might help in visualizing common themes in video transcriptions. These visuals will clearly communicate trends and patterns in the data.

**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, you’ll need to build visualizations such as bar charts, histograms, scatter plots, and heatmaps to explore engagement metrics and their relationships with video attributes. For machine learning, algorithms like classification models (e.g., logistic regression, decision trees) could predict claim\_status, while clustering methods (e.g., K-means) might uncover patterns in engagement. Additionally, text analysis techniques like topic modeling or sentiment analysis can provide insights from video transcriptions.

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

To build the necessary data visualizations, first clean and preprocess the data to ensure accuracy and consistency. Then, use data visualization tools or libraries (e.g., Matplotlib, Seaborn, Tableau) to create charts and plots. Analyze the visualizations to ensure they effectively communicate the insights and adjust as needed based on feedback or new findings.

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

The most applicable variables for visualizations include claim\_status to compare opinions versus claims, verified\_status to analyze engagement differences between verified and non-verified users, and author\_ban\_status to see how permissions affect engagement. Additionally, engagement metrics like video\_view\_count, video\_like\_count, video\_share\_count, and video\_comment\_count are crucial for understanding video popularity and interaction.

* Going back to the Plan stage, how do you plan to deal with the missing data (if any)?
* To deal with missing data, first identify the extent and pattern of the missing values. Depending on the missingness, you can either impute values using techniques like mean, median, or mode for numerical data, and the most frequent value or a placeholder for categorical data, or remove rows/columns with excessive missing values. Assess the impact of these methods on the dataset's integrity and analysis outcomes.

******PACE: Execute Stage**

* What key insights emerged from your EDA and visualizations(s)?
*  **Engagement Patterns**: Verified users’ videos often receive higher engagement (views, likes, shares) compared to non-verified users.
*  **Claim vs. Opinion**: Videos with claims may have different engagement levels compared to videos with opinions, potentially reflecting varying levels of controversy or interest.
*  **Author Status Impact**: Videos from authors with different ban statuses may show distinct patterns in engagement metrics, revealing how community standing affects interaction.
* What business and/or organizational recommendations do you propose based on the visualization(s) built?
*  **Prioritize Verified Users**: Focus on promoting or collaborating with verified users to leverage their higher engagement levels.
*  **Monitor Claims**: Implement additional fact-checking for videos identified as claims to enhance credibility and mitigate misinformation.
*  **Address Author Ban Status**: Consider strategies to engage or rehabilitate authors under scrutiny or banned to potentially improve their content's 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?
*  **Content Trends**: What specific topics or themes in the video transcriptions correlate with higher engagement metrics?
*  **Verification Impact**: How does the verification status of users influence the types of content they produce and its reception?
*  **Ban Status Effects**: What are the long-term effects of author ban status on video engagement and content quality?
*  **Engagement Drivers**: Are there certain features (e.g., video duration, transcription content) that significantly drive higher engagement across different video types?
* How might you share these visualizations with different audiences?
*  **Executive Team**: Use high-level summaries with key insights and impactful visuals (e.g., bar charts, trend lines) to quickly convey the main findings.
*  **Technical Team**: Provide detailed charts and interactive dashboards with drill-down capabilities to explore data nuances and support deeper analysis.
*  **General Public**: Create straightforward, visually appealing infographics and simple charts to make the information accessible and easily understandable.