PACE: Plan Stage

- **What is the overall goal of this project?**

  To prepare TikTok's claim classification dataset for future machine learning modeling by building a dataframe, reviewing column data types, gathering descriptive statistics, and identifying meaningful variables for analysis.

- **Who is your audience for this project?**

  The audience includes the TikTok data science team (technical stakeholders) and cross-functional team members such as operations, finance, and project management leads (non-technical stakeholders).

- **What are the key milestones for this project?**

  ➢ Import dataset and build dataframe

  ➢ Explore and summarize data types and structures

  ➢ Identify and modify relevant variables

  ➢ Gather descriptive statistics

  ➢ Communicate findings and next steps in an executive summary

- **How can you best prepare to understand and organize the provided information?**

  To prepare effectively, I will first carefully review the structure of the dataset, including column names, data types, and descriptions. I'll also explore the project goal and expected outcomes to align my data inspection with what the team needs. Taking notes on which variables might be useful for the machine learning model will help me stay organized and focused during analysis.

- **What questions do you need to answer to complete this project?**

  - ➢ What is the structure of the dataset?
  - ➢ Which columns are relevant or irrelevant?
  - ➢ Are there any missing or unusual values?
  - ➢ What are the ranges and distributions of key variables?
  - ➢ What variables may need to be engineered for the ML model?

- **What follow-along and self-review codebooks will help you perform this work?**

  The example code in the Jupyter notebook provided by Orion will guide my initial approach. In addition, reviewing Python documentation for pandas and NumPy, as well as referencing completed lab examples from earlier in the course, will reinforce my understanding and help troubleshoot any issues. Google Colab and Stack Overflow may also serve as useful resources.

- **What are some additional activities a resourceful learner would perform before starting to code?**

  - ➢ Review similar Python projects to understand best practices
  - ➢ Outline the workflow or tasks in a notebook or mind map
  - ➢ Double-check that the dataset is clean, complete, and ready to work with
  - ➢ Ensure that all necessary Python packages are installed and imported
  - ➢ Clarify with teammates or stakeholders if anything is unclear before beginning the actual coding

- **What are the potential challenges of this project?**

- ➢ Understanding the purpose and meaning of each column
- ➢ Identifying and handling missing or anomalous data
- ➢ Communicating technical results clearly to non-technical team members

PACE: Analyze Stage

- **What kind of data are you working with?**

  The dataset includes 19,383 rows and 12 columns of synthetic TikTok video data including views, likes, shares, download counts, video duration, transcription, claim status, and author account status.

- **What information are you trying to uncover from the data?**

  Data types, missing values, distributions, outliers, and potential relationships between variables such as view count and claim status.

- **How will this data help you achieve your project goals?**

  Analyzing and organizing the data will support the development of a well-prepared dataset for training a predictive model to classify user content as either a claim or an opinion.

- **Will the available information be sufficient to achieve the goal based on your intuition and the analysis of the variables?**

  Yes, the dataset includes a rich combination of numerical and categorical variables, such as `claim_status`, `video_duration_sec`, and `video_view_count`, which can be valuable for building a predictive classification model. The data seems sufficient for initial exploration and model development, especially when combined with additional feature engineering.

- **How would you build summary dataframe statistics and assess the min and max range of the data?**

  I would use the `.describe()` method in pandas to generate summary statistics for each numeric column. This would provide the count, mean, standard deviation, min, and max values. To analyze specific columns, I'd also apply `.min()` and `.max()` individually and visualize distributions using histograms or boxplots.

- **Do the averages of any of the data variables look unusual? Can you describe the interval data?**

  Yes, the average `video_view_count`, `video_like_count`, and `video_share_count` appear to be highly skewed due to the presence of outliers. Some videos likely went viral, causing unusually high values. This makes the mean less reliable than the median for these columns. The data intervals for these variables range from zero to extremely high numbers, suggesting a long-tailed distribution typical of social media metrics.

PACE: Execute Stage

- **How will you share your results and communicate your findings?**

  Via an executive summary document and a completed Python notebook shared with the team.

- **What will be the most important takeaway for your audience?**

  That the dataset has been successfully structured and explored, relevant variables identified, and that it is now ready for further model building and hypothesis testing.

- **How will you ensure your findings are clearly understood?**

By using accessible language in the executive summary, summarizing key variable findings, and recommending specific next steps in the modeling process.

- **Given your current knowledge of the data, what would you initially recommend to your manager to investigate further prior to performing exploratory data analysis?**

  I recommend reviewing the distribution of the `claim_status` values to understand the balance between "claim" and "opinion." We should also investigate how user verification status and ban status relate to claim frequency, which could reveal valuable patterns before deep analysis.

- **What data initially presents as containing anomalies?**

  Columns such as `video_view_count`, `video_like_count`, and `video_download_count` may contain outliers or extremely skewed distributions. We should check for videos with 0 views but high shares or likes, which could indicate logging errors or false data.

- **What additional types of data could strengthen this dataset?**

  Adding a sentiment analysis score for each video transcription, or metadata like publish date/time, user location, or follower count, could enhance our understanding of claim dynamics and improve model performance.