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

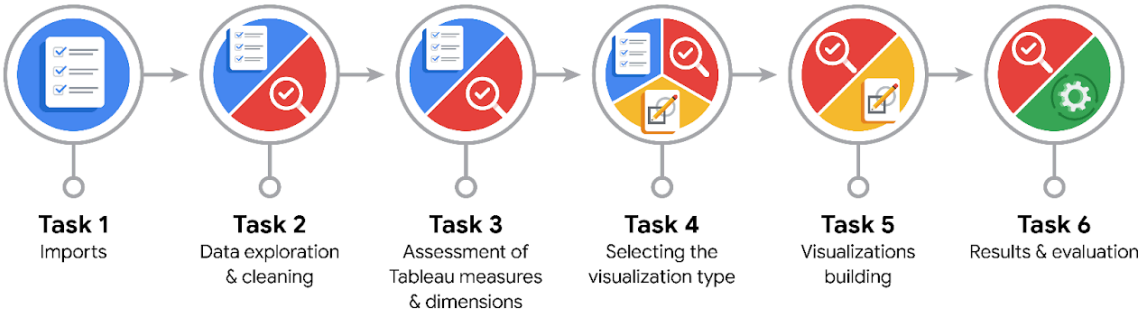
Relevant Interview Questions

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?

verified_status,claim_status,verified_status,author_ban_status,video_like_count,video_view_count,video_share_count,video_download_count,video_comment_count

- What units are your variables in?

String, int

- 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?

Based on the dataset description, I presume that claim videos might have higher engagement (views, likes, or shares) than opinion videos, since claims can be more controversial or attention-grabbing

- Is there any missing or incomplete data?

NO

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

YES

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

Most columns in the dataset appear to be in a consistent format.

Numeric columns such as `video_view_count`, `video_like_count`, and `video_duration_sec` are stored as numbers, while categorical columns like `claim_status`, `author_ban_status`, and `verified_status` are stored as text (object type).

PACE: Analyze Stage

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

- Check data quality (missing values, incorrect data types, duplicates).
- Explore key variables such as `claim_status`, `author_ban_status`, and `verified_status`.
- Summarize numeric features (like views, likes, shares, comments) using descriptive statistics.
- Visualize relationships and distributions using histograms, boxplots, and bar charts.
- Identify correlations and outliers to understand engagement behavior patterns.

- 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.?

No, the given was enough.

Some structuring was needed, such as grouping, filtering, sorting

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

Box plot,bar chart and Histogram

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?

- Visualizations: histograms, boxplots, bar charts, and pie charts for data distribution and comparison.
- ML Algorithms: a classification model (e.g., Logistic Regression or Decision Tree) to predict whether a video is a *claim* or *opinion*.
- Outputs: summary statistics and model accuracy metrics.

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

Clean and preprocess the dataset (handle missing or inconsistent data).

Use grouping (`groupby()`) and aggregation (`agg()`) for summarization.

Generate visualizations using **Matplotlib** or **Seaborn**.

Add labels, titles, and color schemes for clarity and better presentation.

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

1. `claim_status`
2. `author_ban_status`
3. `verified_status`
4. Video_view_count
5. `video_like_count`
6. `video_share_count`
7. `video_duration_sec`

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

  - For numeric data → fill with mean/median values.
  - For categorical data → fill with mode or "Unknown."
  - Drop rows if missing values are excessive and uninformative

PACE: Execute Stage

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

- The dataset contains more *claim* videos than *opinions*.
- *Claim* videos tend to have higher average views and shares than *opinion* videos.
- Verified users' videos generally receive more engagement.
- Some outliers exist — a few videos have extremely high view counts compared to others.

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

- The dataset contains more *claim* videos than *opinions*.
- *Claim* videos tend to have higher average views and shares than *opinion* videos.
- Verified users' videos generally receive more engagement.
- Some outliers exist — a few videos have extremely high view counts compared to others.

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

Does video duration influence engagement (views, likes, shares)?

Do banned or scrutinized authors create more claim-based videos?

How does verified status correlate with misinformation risk?

- How might you share these visualizations with different audiences?

- **For TikTok executives:** concise dashboard or PowerPoint with summarized insights.

- **For technical analysts:** Jupyter Notebook with full code and graphs.
- **For public reports:** visual summaries or infographics highlighting the main findings.