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~~

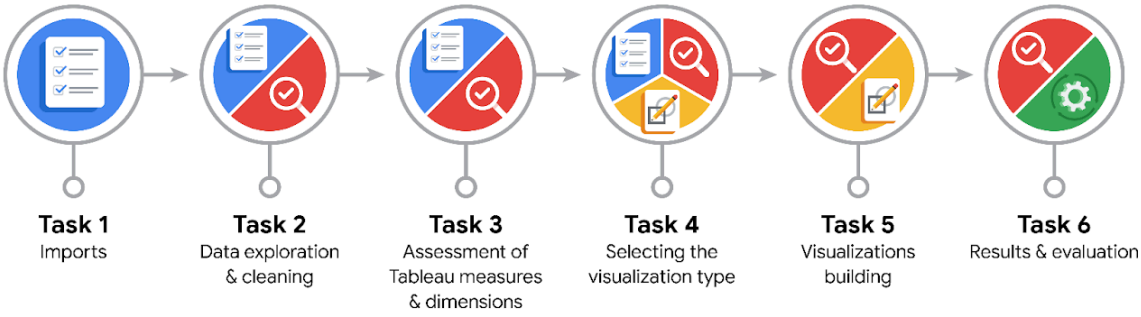
Relevant Interview Questions

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?

The dataset contains columns like claim_status, video_id, video_duration_sec, video_view_count, video_like_count, author_ban_status, and video_transcription_text. For this project, the most relevant columns are claim_status, engagement metrics (views, likes, shares), and author_ban_status since they relate directly to content classification and engagement analysis.

- What units are your variables in?

Most variables are counts (views, likes, shares, comments, downloads), while `video_duration_sec` is in seconds, and categorical variables such as `claim_status` or `author_ban_status` are string labels.

- 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?

Videos labeled as claims may have higher engagement metrics than opinions, and banned authors may skew engagement rates. Outliers are expected in views and likes due to viral content.

- Is there any missing or incomplete data?

Yes, some variables like claim_status, video_transcription_text, and engagement metrics have missing values that must be addressed during cleaning.

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

No, the dataset has mixed formats: numerical, categorical, and text columns, requiring careful handling in analysis and visualization.

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

Data inspection (info(), describe()), handling missing values, calculating summary statistics, exploring distributions, and initial visualizations like histograms and boxplots will be necessary.

PACE: Analyze Stage

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

Steps include inspecting distributions of numeric variables, assessing relationships between claim_status and engagement metrics, calculating summary statistics, and identifying potential outliers or anomalies.

- 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.?

Joining additional datasets is not required for this project. Structuring steps include filtering by claim_status or author_ban_status, sorting by engagement metrics, and creating new computed columns like likes_per_view or shares_per_view.

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

Bar charts for categorical variables, boxplots for engagement metrics, and scatter plots to show relationships between video metrics will communicate insights effectively. Heatmaps or grouped bar charts may help visualize differences across claim_status and author_ban_status.

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 include histograms, boxplots, scatter plots, and bar charts. Machine learning algorithms may include a classification model (logistic regression or decision tree) to predict claim vs opinion.

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

Processes include cleaning missing values, computing new engagement rate columns, grouping data by categorical variables, and summarizing statistics for visualization.

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

Key variables are claim_status, author_ban_status, video_view_count, video_like_count, video_share_count, and computed engagement rate metrics like likes_per_view.

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

Missing numeric values can be filled with median or mean imputation, and missing categorical variables can be labeled as "Unknown". Certain rows may be dropped if missing critical identifiers like claim_status.

PACE: Execute Stage

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

  Videos labeled as claims tend to receive significantly higher views, likes, and shares compared to opinions. Banned authors and those under review often have higher engagement, suggesting stricter moderation impacts visibility and engagement

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

  Focus moderation efforts on high-engagement claim videos to prevent misinformation spread. Engagement patterns by author status could inform automated prioritization of user reports.

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

  Investigate correlations between video length and engagement, the impact of verified authors on engagement, and trends in specific content topics. Analyze temporal trends to see if engagement changes over time.

- How might you share these visualizations with different audiences?

  Interactive dashboards (Tableau, Power BI) for stakeholders, simplified static charts for management presentations, and annotated visualizations in reports for technical team members.