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

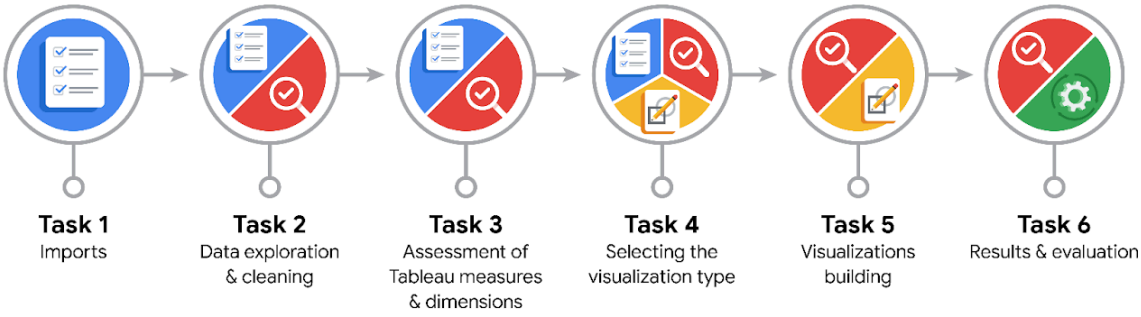
Relevant Interview Questions

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?

Columns include: #, `claim_status`, `video_id`, `video_duration_sec`, `video_transcription_text`, `verified_status`, `author_ban_status`, `video_view_count`, `video_like_count`, `video_share_count`, `video_download_count`, `video_comment_count`.

Most relevant for this deliverable: `claim_status`, `video_duration_sec`, `verified_status`, `author_ban_status`, `video_view_count`, `video_like_count`, `video_comment_count`, `video_share_count`, and `video_download_count`. These are the most relevant because they will help determine the difference between claim and opinion videos.

- What units are your variables in?

`video_duration_sec`: seconds.
`video_view_count`, `video_like_count`, `video_share_count`, `video_download_count`, `video_comment_count`: numerical counts.
`claim_status`, `verified_status`, `author_ban_status`: categorical.
`video_id` and #: ID numbers.

- 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?

I expect that videos with higher engagement metrics (views, likes, comments, shares, downloads) will likely be claim videos. I presume that banned authors will have a higher proportion of claim videos. I expect to see right skewness in the engagement counts.

- Is there any missing or incomplete data?

Yes, `claim_status`, `video_transcription_text`, `video_view_count`, `video_like_count`, `video_share_count`, `video_download_count` , and `video_comment_count` have missing values.

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

No, there are numerical (int, float) and categorical (object) data types.

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

Descriptive statistics, data visualization (histograms, boxplots, scatter plots, bar charts, pie charts), handling missing data, and outlier detection.

PACE: Analyze Stage

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

1. Handle missing data (impute or remove). 2. Calculate descriptive statistics to understand the data's central tendencies and spread. 3. Visualize distributions of numerical variables. 4. Explore relationships between variables using scatter plots and bar charts. 5. Analyze categorical variables and their impact on engagement metrics. 6. Identify outliers and assess their potential impact.

- 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, joining is not required for this analysis. Structuring: Filter data to analyze specific subsets (e.g., verified vs. unverified users), sort data to identify trends, and group data to calculate aggregated metrics.

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

• Bar charts and pie charts for categorical data. • Histograms and boxplots for numerical distributions. • Scatter plots for relationships. • I plan to keep the visuals simple and clear, with appropriate labels and titles, for easy understanding by a non-data-savvy audience, and use high contrast colors for accessibility.

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?

Histograms, boxplots, bar charts, pie charts, and scatter plots.

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

• Use `matplotlib` and `seaborn` libraries in Python to create the visualizations. • Clean and preprocess the data as needed. • Customize visualizations with appropriate titles, labels, and colors.

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

`claim_status`, `video_duration_sec`, `verified_status`, `author_ban_status`, `video_view_count`, `video_like_count`, `video_comment_count`, `video_share_count`, and `video_download_count`.

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

For numerical columns, I will impute with the median due to the data's skewness. For `claim_status` and `video_transcription_text`, I will remove the rows with missing values, as these are important to the analysis.

PACE: Execute Stage

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

• Claim videos have significantly higher view counts than opinion videos. • Non-active authors (banned, under review) tend to post more claim videos and have higher view counts. • Engagement metrics are highly right-skewed. • Verified users are more likely to post opinions.

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

• Implement stricter content review processes for claim videos to ensure accuracy and prevent misinformation. • Investigate the reasons behind the higher engagement of claim videos and use those insights to improve content strategy. • Provide more resources to unverified accounts, as they are the large majority of accounts.

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

• What are the specific topics or themes that contribute to the virality of claim videos? • How does video duration affect engagement metrics? • What are the key differences in the transcription text between claim and opinion videos?

- How might you share these visualizations with different audiences?

• For technical audiences: Share the Jupyter notebook and detailed reports. • For non-technical audiences: Use Tableau dashboards with clear and concise visualizations and explanations. • For visually impaired audiences: Ensure high contrast, use descriptive titles and labels, and provide alternative text descriptions for all visualizations.