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

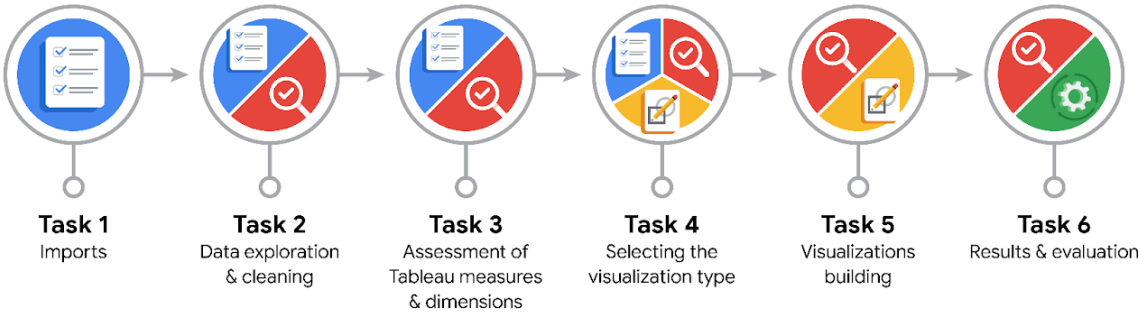
Relevant Interview Questions

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?

  Video comment count, video like count, video share count, video download count, video view count, claim status, author ban status

- What units are your variables in?

  int, float, obj

- 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?

  Video content popularity (number of views, downloads etc.) is more for claim videos, and it is more likely that the content author is/was banned from the platform.

- Is there any missing or incomplete data?

  From the result of the describe() method, video view / like / download / share / comment count columns have some missing data as the total count value is less.

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

There are varying data types in the columns.

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

Discovering, Structuring, Validating, Presenting

PACE: Analyze Stage

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

Understand the dataset using various pandas methods like info() and describe(), use plotting packages in Python to plot variables and observe their trends while using easy to understand colour schemes.

- 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, this dataset has a lot of rows and not many missing values and so is sufficient to carry out EDA. The missing rows need to be removed before further analysis is carried out.

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

To understand outliers, box plots and histograms will be best suited. To understand the trends between variables, scatter or line plots will be appropriate.

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?

Box plots, histograms, scatter plots.

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

Remove missing value rows first.

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

Video_view_count, video_like_count, video_share_count, video_download_count, video_comment_count.

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

If the number of missing data is low enough, just do not use them in the analysis. If there are a large number of missing rows, it needs to be addressed in a manner that is appropriate such as filling in with the average.

PACE: Execute Stage

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

Viewer engagement is much more for claim videos than for opinion videos. Also, banned authors are much more likely to have posted claim videos compared to active authors.

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

For the prediction model, the variables that represent viewer engagement with the videos should be considered as they seem to carry the most predictive information regarding whether a video is a claim or an opinion.

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

EDA can find out whether banned / active / verified / unverified users are more likely to post claim vs opinion videos.

- How might you share these visualizations with different audiences?

Using appropriate colour schemes to highlight the message in the graph or chart. Also, use accessible formats in the visualizations.