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

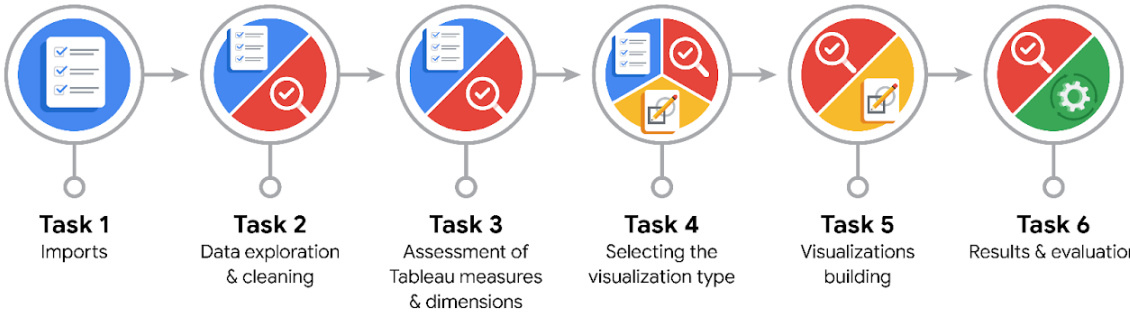
Relevant Interview Questions

Completing the end-of-course project will help you respond 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?

There are 12 columns and the variables that are most relevant to my deliverable are claim status, video duration sec, video like count, video view count, video comment count, video share count, video download count and author ban status.

- What units are your variables in?

Floats, Strings and integers.

- 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?

That the trend will be that videos with banned or under scrutiny status will have higher video view, like, share, comment and download counts.

- Is there any missing or incomplete data?

Yes there are several null values.

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

Yes.

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

Discovering and Structuring

PACE: Analyze Stage

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

To follow the 6 steps of EDA with validating after each modification to the dataset.

- 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 type of structuring that needs to be done to this dataset is filtering.

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

Boxplots, Histograms, Scatter Plots, Bar Graphs and Tableau Dashboards.

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?

Boxplots, Histograms, Scatter Plots, Bar Graphs and Tableau Dashboards.

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

The 6 processes of EDA

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

Claim status, Author ban status, video like, view, share, download and comment counts.

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

I would see the importance of the missing data and then decide on whether to delete, inform the owner or replace the missing data with average/median values from available data.

PACE: Execute Stage

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

That the vast majority of video views come from content labeled as claims rather than opinions. The median view counts are significantly higher for authors who are either under review or banned. Visualizations like scatter plots and box plots revealed a strong positive correlation between video views and likes and the distribution of those variables is highly right skewed, requiring special outlier handling using the median + 1.5 x IQR method.

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

Monitor claim content more closely, as it drives most views. Flag high – engagement posts from banned or reviewed accounts. Promote verified opinion content to reduce misinformation.

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

I could explore whether certain video topics or hashtags are more likely to be flagged or go viral? And whether video duration influences engagement differently for claims versus opinions.

- How might you share these visualizations with different audiences?

I would use simplified dashboards with tooltips and color coding for non-technical audiences, and include detailed filters, legends and data labels for analysts or internal stakeholders.