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

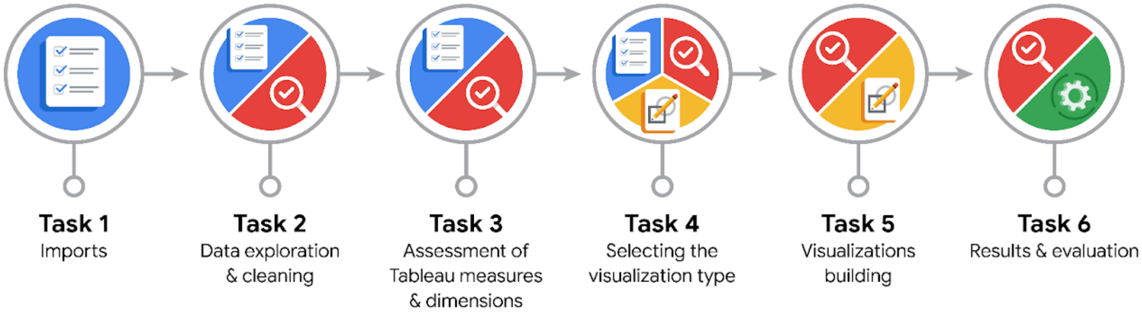
Relevant Interview Questions

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.


Task 1
Imports
Task 2
Data exploration & cleaning
Task 3
Assessment of Tableau measures & dimensions
Task 4
Selecting the visualization type
Task 5
Visualizations building
Task 6
Results & evaluation

Data Project Questions & Considerations

PACE: Plan Stage

- What are the data columns and variables and which ones are most relevant to your deliverable?

The engagement values and the video duration alongside the author ban status help understand the difference behind the things that lead to the final claim status.

- What units are your variables in?

Duration in seconds, view count in million, everything else in 1s

- 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?

Engagement statistics could inform if something is a claim vs opinion, engagement data has a long right tail. Authors status might help to see if something is a claim or opinion.

- Is there any missing or incomplete data?

There are missing values in many columns.

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

Engagement stats are floats, all other numericals are integers with the rest being strings.

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

Managing missing values, cleaning the data, understand structure and document observations and hypotheses.

PACE: Analyze Stage

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

Know the problem and the objective, understand and get familiar with the dataset, clean the dataset and add any necessary information through joins or merges, create visualisations of findings and present the data in the way it is found without any bias.

- 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.?

Currently no additional data is needed but when adding data, it needs to be in the same format as the current data and it needs to be filtered so no type errors or additional missing values are added.

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

Box plots and histograms for distributions and outliers and scatter graphs to show trends.

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?

Visualitions of the data, machine learning model to predict claims and visualisations of its efficiency and report of findings and maybe different models if they have different caveats e.g. high precision/recall.

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

Data cleaning and filtering to prepare for building visualizations, then aggregations and grouping to create different graphs.

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

Engagement metrics + duration as they are numeric values.

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

Depending on the column, either use an imputer or drop them if there is not many or have other erroneous values alongside them.

PACE: Execute Stage

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

Most data has a long right tail with the expectation of video duration which has some peaks at 4 different length marks, most opinion claims are for lower viewed videos while claim claims are for all popularities of videos.

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

Look into the more popular videos and check if opinion claims are being ignored or always rejected.

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

Why is the opinion claim status only present in lower viewed videos?

- How might you share these visualizations with different audiences?

Change the coloring or wording, change which visualizations are shown based on the requirements of the audience