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

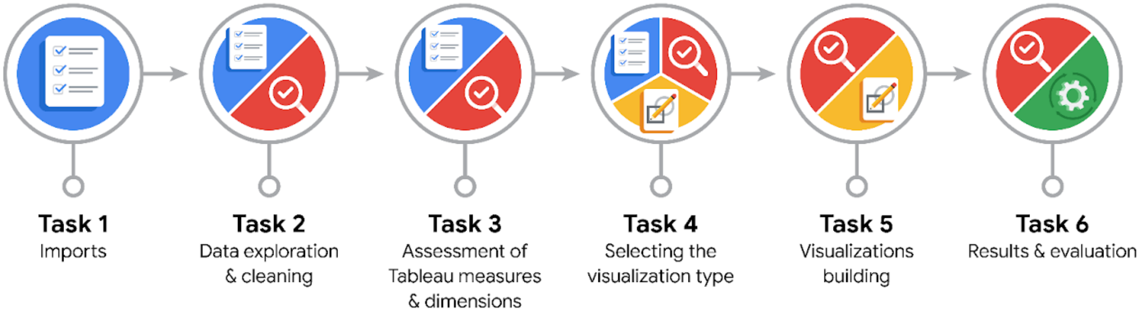
Relevant Interview Questions

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.


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?

Columns are about engagement metric and all are relevant to analyze for an outcome.

- What units are your variables in?

float64(5), int64(3), object(4).

- 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?

Data exploration and cleaning.

- Is there any missing or incomplete data?

No.

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

float64(5), int64(3), object(4).

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

Discovering which can check out the overall shape, size, and content of the dataset. You find it is short on data.

PACE: Analyze Stage

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

Cleaning which can check for outliers, missing data, and needs for conversions or transformations.

- 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.?

It depends on locations where data is kept in different files.

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

The visualizations most helpful for considering the distribution of the data include box plots and histograms. Visualizing the distribution of the data can inform the next steps and considerations in data analysis. For example, data distribution will inform which types of modeling is needed.

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 and histograms.

For loop,

Statistics matric.

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

Import packages for data visualization.

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

Video view count and video like count.

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

– Delete them: If you are sure the outliers are mistakes, typos, or errors and the dataset will be used for modeling or machine learning, then you are more likely to decide to delete outliers. Of the three choices, you'll use this one the least.

– Reassign them: If the dataset is small and/or the data will be used for modeling or machine learning, you are more likely to choose a path of deriving new values to replace the outlier values.

– Leave them: For a dataset that you plan to do EDA/analysis on and nothing else, or for a dataset you are preparing for a model that is resistant to outliers, it is most likely that you are going to leave them in.

PACE: Execute Stage

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

Bar chart of claim and opinion.

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

Online entertainment industry

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

I want to further investigate distinctive characteristics that apply only to claims or only to opinions. Also, I want to consider other variables that might be helpful in understanding the data.

- How might you share these visualizations with different audiences?

Put yourself in your client's perspective, what would they want to know?