Course One
Foundations of Data Science

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 1 PACE strategy document
- ☐ Answer the questions in the Jupyter notebook project file
- ☐ Complete coding prep work on project's Jupyter notebook
- ☐ Summarize the column Dtypes
- ☐ Communicate important findings in the form of an executive summary

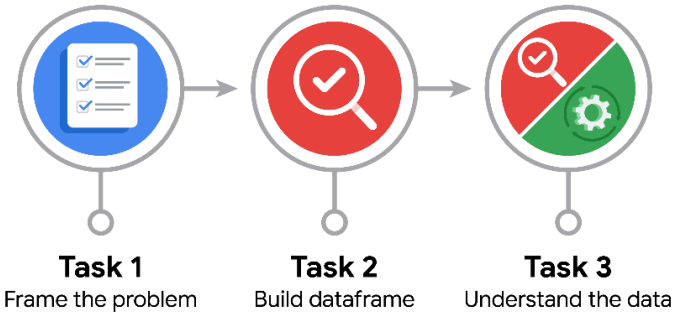
Relevant Interview Questions

Completing the end-of-course project will help you respond these types of questions that are often asked during the interview process:

- Describe the steps you would take to clean and transform an unstructured data set.
- What specific things might you look for as part of your cleaning process?
- What are some of the outliers, anomalies, or unusual things you might look for in the data cleaning process that might impact analyses or ability to create insights?

Reference Guide

This project has three tasks; the visual below identifies how the stages of PACE are incorporated across those tasks.



Data Project Questions & Considerations

PACE: Plan Stage

- How can you best prepare to understand and organize the provided information?

To best prepare, I need to understand the stakeholders' goals and review the dataset structure, variables, and data types. This will help me organize the information effectively and ensure the data supports the project objectives.

- What follow-along and self-review codebooks will help you perform this work?

I will refer to the dataset documentation, including the data dictionary and column descriptions, to understand the structure of the data and the meaning of each variable.

- What are some additional activities a resourceful learner would perform before starting to code?

Before starting to code, a resourceful learner would review the project scenario, understand the stakeholders' goals, study the data dictionary, and learn about the company's structure and business context to ensure the analysis aligns with real-world needs.

PACE: Analyze Stage

- Will the available information be sufficient to achieve the goal based on your intuition and the analysis of the variables?

Yes, the available information is sufficient. The dataset provides clear numerical and categorical indicators that can help a machine learning model recognize patterns in video characteristics. These patterns support faster and more accurate classification of videos, which aligns with stakeholder goals of improving content review efficiency.

- How would you build summary dataframe statistics and assess the min and max range of the data?

I would use pandas functions such as `describe()` to review overall summary statistics and `agg()` to calculate specific measures like mean, median, minimum, and maximum values. This helps identify the data range, detect outliers, and understand whether the variables fall within realistic intervals.

- Do the averages of any of the data variables look unusual? Can you describe the interval data?

Some average values appear notably different between claim and opinion videos. For example, engagement metrics such as views, likes, and shares tend to be significantly higher for claim videos. This may reflect differences in user interest or content style. The interval ranges of some engagement variables also suggest the presence of outliers, possibly due to viral or trending videos.

PACE: Construct Stage

Note: The Construct stage does not apply to this workflow. The PACE framework can be adapted to fit the specific requirements of any project.

PACE: Execute Stage

- Given your current knowledge of the data, what would you initially recommend to your manager to investigate further prior to performing exploratory data analysis?

I would recommend investigating potential outliers and extreme values in engagement metrics such as views, likes, and shares. These unusually high or low values may distort the analysis and should be verified before moving forward with deeper exploratory data analysis.

- What data initially presents as containing anomalies?

Engagement-related variables, particularly video view count, like count, and share count, show signs of anomalies. Some videos have extremely high engagement compared to the average, which suggests the presence of viral or trending content that may act as outliers.

- What additional types of data could strengthen this dataset?

Additional contextual data would strengthen the dataset, such as video descriptions, hashtags, upload time, and video duration. These features could provide deeper insight into why certain videos receive higher engagement and help improve the performance of the classification model.