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 PACE Strategy Document to plan your project while considering your audience members, teammates, key milestones, and overall project goal.
- Create a project proposal for the data team.

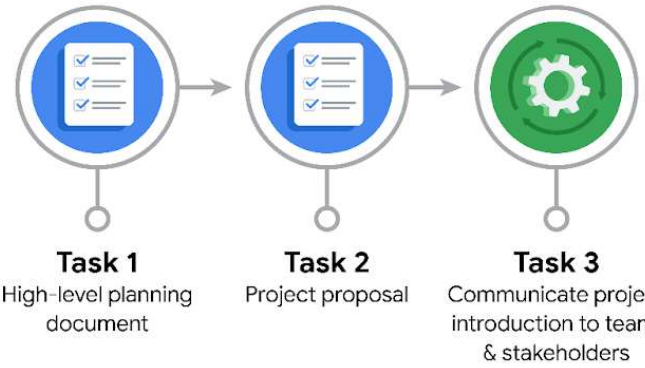
Relevant Interview Questions

Completing this end-of-course project will empower you to respond to the following interview topics:

- As a new member of a data analytics team, what steps could you take to get 'up to speed' with a current project? What steps would you take? Who would you like to meet with?
- How would you plan an analytics project?
- What steps would you take to translate a business question to an analytical solution?
- Why is actively managing data an important part of a data analytics team's responsibilities?
- What are some considerations you might need to be mindful of when reporting results?

Reference Guide

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



Data Project Questions & Considerations

PACE: Plan Stage

- Who is your audience for this project?

The target audience varies between technical and non-technical, and the communication to each must be formulated appropriately. As part of the Data Team, we have Willow Jaffey, Rosie Mae Bradshaw, and Orion Rainier. Cross-functional team members consist of Mary Joanna Rodgers, Margery Adebowale, and Maika Abadi.

- What are you trying to solve or accomplish? And, what do you anticipate the impact of this work will be on the larger needs of the client?

The goal of the project is to develop a machine learning model that can classify whether a video provides a claim or offers an opinion. This model will help alleviate some of the workload of the moderators who need to manually vet these claims.

- What questions need to be asked or answered?

1. What is the data source?
2. Was the data acquired in a manner that is representative of the population at large?
3. How long is the project approximately going to take?
4. Are there any ethical concerns with any aspects of the model being developed?

5. What is the communication hierarchy going to look like? Who needs to be directly updated on the project's during each milestone?

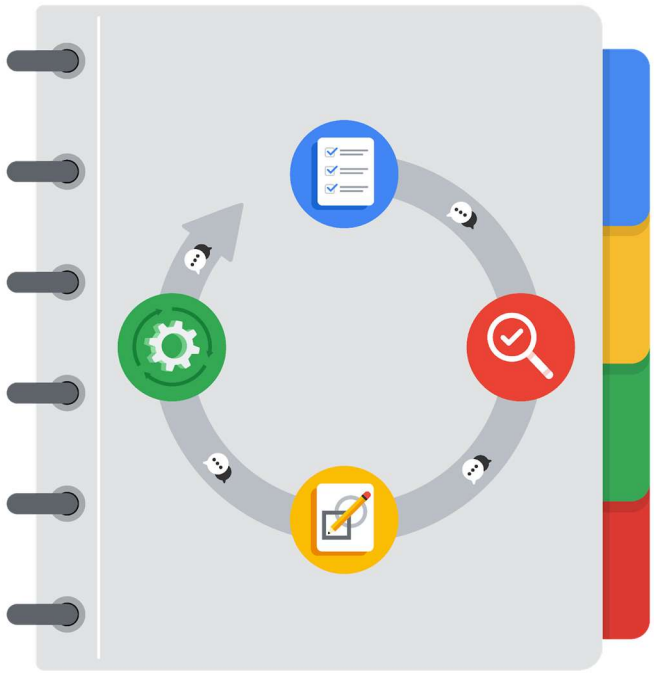
- What resources are required to complete this project?

For one, access to the data and information about its collection methods is crucial in case questions arise or additional data is needed. Additionally, the various Python packages necessary for basic exploratory data analysis (EDA) include NumPy, Pandas, Matplotlib, Seaborn, and Scikit-learn will be required.

- What are the deliverables that will need to be created over the course of this project?

For this section, a Project Proposal will be created, outlining the overall project and its deliverables. For subsequent sections, a PACE Strategy Document and Executive Summary will be created for stakeholders.

THE PACE WORKFLOW



[Alt-text: The PACE Workflow with the four stages in a circle: plan, analyze, construct, and execute.]

You have been asked to demonstrate for the company's data team how you would use the PACE workflow to organize and classify tasks for the upcoming project. Select a PACE stage from the dropdown buttons. A few tasks involve more than one stage of the PACE workflow. Additionally, not every workplace scenario will require every task. Refer back to the Course 1 end-of-course portfolio project overview reading if you need more information about the tasks within the project.

Project tasks

Following are a group of tasks your company's data team has determined need to be completed within this project. The data analysis manager has asked you to organize these tasks in preparation for the project proposal document. First, identify which stage of the PACE workflow each task would best fit under using the drop down menu. Next, give an explanation of why you selected the stage for each task. Review the following readings to help guide your selections and explanation: The PACE stages and Communicate objectives with a project proposal. You will later reorder these tasks within a project proposal.

1. **Evaluating the model:** Construct

   Why did you select this stage for this task?

At this stage, any models developed are built, evaluated and revised as needed.

2. **Conduct hypothesis testing:** Analyze

   Why did you select these stages for this task?

At this stage, data is examined for patterns, relationships and correlations with one another and the target variable. Hypothesis testing is then conducted to determine whether the original hypothesis is supported.

3. **Begin exploring the data:** Analyze

   Why did you select this stage for this task?

During the EDA phase, data exploration occurs, which falls under the Analyze stage of the PACE framework.

4. **Data exploration and cleaning:** Analyze

   Why did you select these stages for this task?

Cleaning the data, normalization of values and exploration occur during the EDA stage, which ultimately falls under the Analyze stage of the PACE framework.

5. **Establish structure for project workflow (PACE):** Plan

   Why did you select this stage for this task?

During the Plan section, the overall workflow and deliverables are established. Adjustments are made throughout, but the major milestones should be oulined.

6. **Communicate final insights with stakeholders:** Execute

   Why did you select this stage for this task?

Final insights are communicated to stakeholders during the Execute stage. Once the data has been cleaned, hypothesis testing has been completed and models have been built and revised, the results of these findings are compiled and communicated appropriately to their respective audiences.

7. **Compute descriptive statistics:** Analyze

   Why did you select this stage for this task?

Prior to model development, descriptive statistics is completed in the Analyze phase during the initial investigation. Key features of the data are considered, such as central tendency (mean, median and mode), measures of variability (standard deviation, variance and range), frequency distributions (frequency tables, histograms and visualizations (bar charts, line charts and scatter plots.)

8. **Visualization building:** Analyze **and** Execute

   Why did you select these stages for this task?

Visualizations are built in both the Analyze and Execute stages, when data is first investigated and when final insights are communicated to stakeholders.

9. **Write a project proposal:** Plan

   Why did you select this stage for this task?

The initial outline of the project, the Project Proposal, is completed during the Plan phase. This provides an overall roadmap of the major milestones required during the course of the project as a whole.

10. **Build a regression model:** Analyze **and** Construct

    Why did you select this stage for this task?

During the Analyze phase, when variable relationships are tested and investigated, regression models are typically built at this time to assist with testing hypothesis. That being said, these models will often be useful during the Construct phase, when machine learning models are built, tested and refined.

11. Compile summary information about the data: Plan

Why did you select this stage for this task?

During the Plan phase, a higher-level summary of the information is often compiled along with information such as a data dictionary or metadata.

12. Build machine learning model: Construct

Why did you select this stage for this task?

During the Construct phase, models are built, validated, compared and refined to other models to find the highest performing one. Once the champion model is found, it can be tested on the test set to acquire final evaluation metrics.