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.

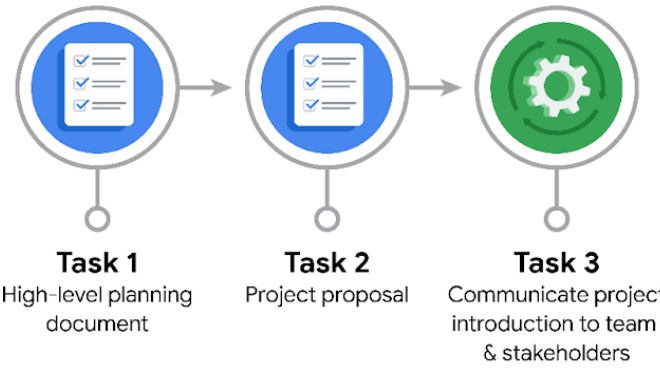
Relevant Interview Questions

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 audience for the project is the New York Taxi & Limousine Commission (TLC), a company that wants to develop a regression model that predicts a taxi cab fare before the ride occurs.

- 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?

I am crafting a project proposal to set up milestones and deliverables for project stakeholders the TLC. A clear proposal will streamline the project workflow.

- What questions need to be asked or answered?

Are the data samples representative of the population? What variables in the dataset are the most relevant features to the target? How much data is TLC providing to train the model? How can I reduce bias in the sample data?

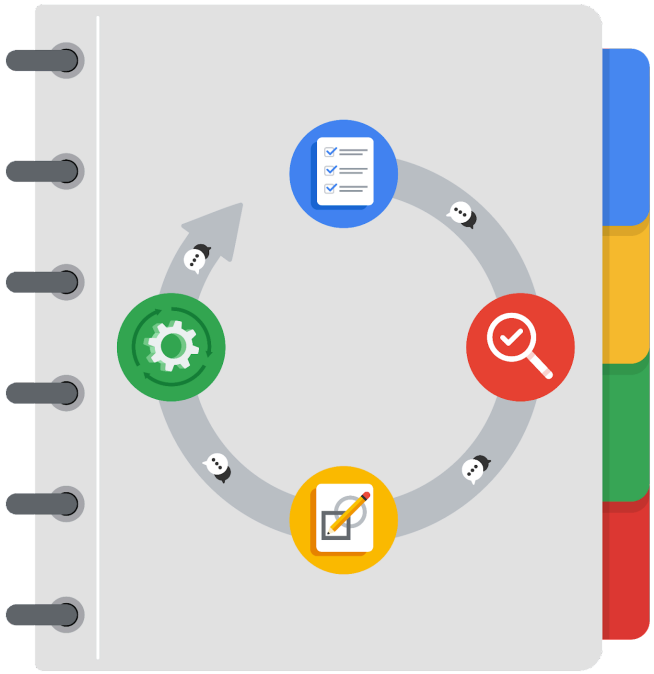
- What resources are required to complete this project?

To complete the project, a data infrastructure is necessary to store the TLC data and regression model. Python .ipynb files will be used for analysis and model development. Feedback and new data to supplement insights discovered by the current data may be needed to improve model performance.

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

Data cleaning to prepare the dataset. Data analysis to discover relationships between variables. Generating visualizations to illustrate findings. Statistical analysis to determine statistical significance. Regression model building for model selection. Updating stakeholders about project progression and potential setbacks.

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:** Execute

   Why did you select this stage for this task?

After the model is built, its performance must be reviewed to determine if it meets stakeholder expectations. The model may be further fine-tuned or more data is added to improve its performance.

2. **Conduct hypothesis testing:** Analyze **and** Construct

   Why did you select these stages for this task?

For the analyze stage, hypothesis testing should be performed when exploring the data. The conclusions drawn from the data need to have a significant difference. Hypothesis tests are developed in the construct phase when declaring null and alternative hypotheses.

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

   Why did you select this stage for this task?

Data exploration is a process where the data must be preprocessed (data cleaning, data wrangling) to see variable relationships, verify data types, and ensure empirical results.

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

   Why did you select these stages for this task?

Data exploration is a process where the data must be preprocessed (data cleaning, data wrangling) to see variable relationships, verify data types, and ensure empirical results. The steps preprocessing should be defined ahead of time based on the size of the dataset and column types/cardinality.

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

   Why did you select this stage for this task?

The project workflow should be set before any work is performed on the data. This is done so realistic deadlines are set, team members are aware of their responsibilities, and setbacks do not largely impact the project schedule.

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

   Why did you select this stage for this task?

Any recommendations or final observations shared with stakeholders always come at the final phase of a project cycle.

7. **Compute descriptive statistics:** Analyze

   Why did you select this stage for this task?

Any computations on the data is a form of data analysis to drive insights on relationships the client wants revealed.

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

   Why did you select these stages for this task?

The analyze stage was selected as visualizations are generated during analysis to show relationships between variables and distributions. The execute stage was selected as visualization aid non-technical stakeholders when interacting with trends discovered in the data, such as interactive dashboard showing ride length vs trip duration.

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

   Why did you select this stage for this task?

The project proposal sets milestones to be completed for the project and is an overview of the project cycle.

10. Build a regression model: Construct and Analyze

Why did you select this stage for this task?

Building a regression model is where the steps from preprocessing the data take form so that predictions on the data can be formed. The data must be preprocessed during the analyze stage to handle categorical data types, missing values, and outliers.

11. Compile summary information about the data: Analyze

Why did you select this stage for this task?

Summarizing statistics about the data is a part of preprocessing. Summaries reveal outliers, unintentional data appearances, and show what most of the sample is consistent to.

12. Build machine learning model: Construct

Why did you select this stage for this task?

The machine learning model takes multiple approaches to its development, which occurs in the construct phase, to ensure accurate predictions on the data.