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.~~

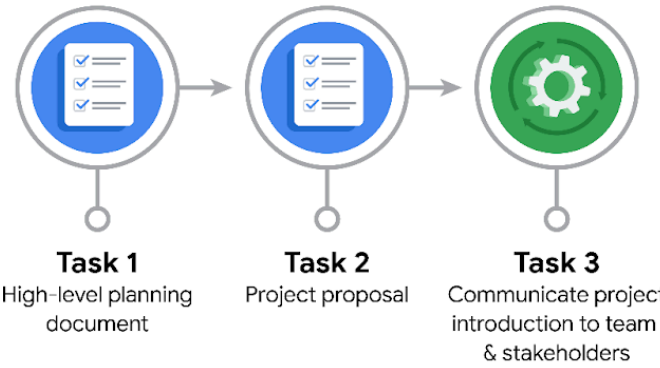
Relevant Interview Questions

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?

Tiktok Data Team:

Willow Jaffey- Data Science Lead

Rosie Mae Bradshaw- Data Science Manager- also my supervisor.

Orion Rainier- Data Scientist

Tiktok co-workers outside the data team:

Mary Joanna Rodgers- Project Management Officer

Margery Adebowale- Finance Lead, Americas

Maika Abadi- Operations Lead

- 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?

What we are trying to solve is the issue of acknowledging a high number of user reports against TikTok content to classify it as either an opinion or a claim. By developing a predictive model to determine whether a video contains a claim or offers an opinion, TikTok aims to automate this classification process. This approach will help reduce the backlog in user reports, allowing them to be prioritized more efficiently by the TikTok moderators.

The impact would be better content moderation across videos, user comments, and content claims, ultimately ensuring higher-quality content, reducing the spread of misinformation, and improving the overall user experience on the platform.

- What questions need to be asked or answered?

- What is the source of data, and will it be only from primary sources or will secondary sources be considered?
- How will the missing data be acknowledged and dealt with?
- What data will be the most useful data in TikTok's database for this project?
- What statistical tests will be done?
- What type of regression model will be appropriate for this project?
- What is the condition of the provided dataset?
- What variables will be the most useful?
- Are there trends within the data that can provide insight?
- What steps can be taken to reduce the impact of bias?
- What will be the main talking points for us going forward into the execute phase?
- What hypothesis testing method will be used for the dataset?
- What are the assumptions made by any regression model developed for this project?

- What resources are required to complete this project?

The resources needed to complete this project are:

- The data from primary and secondary sources, which will be preprocessed and processed for further analysis and other tasks like training, testing, and validation.
- The Python notebook to implement the analysis, model development, and evaluation.
- Input from stakeholders to ensure the relevance and context of the data and analysis.
- The software and hardware resources required for this project, including computational resources for model training and validation.

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

The deliverables needed to created are:

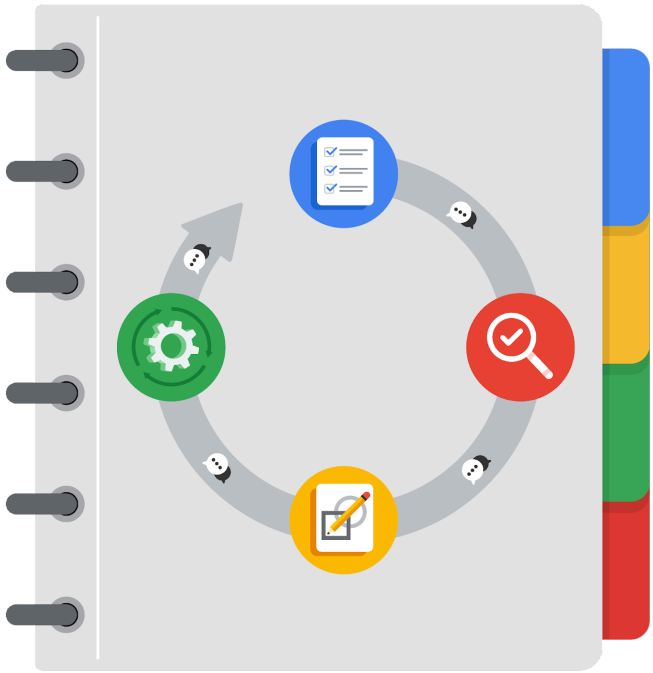
-Project Proposal

-The database being ready for training/modelling, scrubbed for exploratory data analysis

-Stakeholders being updated at every stage of the PACE framework or in accordance with milestone completion

-Machine learning model and algorithm being developed -Acquiring the trained machine learning model -Visualizations for insights and data understanding -Executive summary -Statistical model, regression analysis, and/or machine learning model outcomes

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?

   As evaluation is a subsequent process done after developing the model, I selected the Construct phase for it, as it comes under this phase. Evaluation includes validating and testing the model against new unseen data to ensure it meets the project's expectations and goals. This process checks if the model performs as expected on real-world data and is capable of delivering the desired outcomes.

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

   Why did you select these stages for this task?

   I selected the Analyze phase as the hypothesis formation and hypothesis testing come under the processing and analysis step/phase, where, after the hypothesis is formed, it is then analyzed based on data and statistical methods. Additionally, I also selected the Construct phase, as with model formation and evaluation, it will also help us conduct hypothesis testing using the advanced model and statistical methods. During this phase, the statistical tests are carried out to validate the hypothesis and assess the model's performance on real-world data.

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

   Why did you select this stage for this task?

   It's the Analyze phase where the data is preprocessed, processed, and, with the insights gained, the data and the acquired insights are further analyzed to infer actionable insights and catalyze business decisions. This is the phase that includes EDA (Exploratory Data Analysis), which entails exploring the existing data to infer patterns, trends, and other key features, helping to deepen the understanding of the dataset and the information within it. This phase plays a crucial role in identifying important variables and uncovering trends that can guide further analysis and decision-making.

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

   Why did you select these stages for this task?

   Data exploration and cleaning take place in the Analyze phase, as in this phase we perform preprocessing (cleaning, formatting, dealing with missing values, etc.) and exploratory analysis (inferring the structure, quality, and key characteristics) of the data. This phase is critical for gaining a deeper understanding of the dataset and ensuring that the data is in a suitable form for further

analysis or model development. It helps identify issues like missing or inconsistent values and provides insights into the patterns and relationships within the data.

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

Why did you select this stage for this task?

TThis is done in the Plan phase, as this is the phase where the scaffold or the blueprint of the project is formulated and the roles are established (a RACI chart may be used). The project workflow is formed (PACE workflow may be used). Additionally, the Analyze phase can particularly incorporate the APPASA framework, and the same can be applied globally to the whole project's workflow. The Plan phase sets the foundation for the project's execution by defining goals, setting priorities, and organizing the work structure, ensuring clarity and alignment throughout the project.

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

Why did you select this stage for this task?

It's the Execute phase where the final results, developed model, inferred insights, and visualizations are shared with the stakeholders to convey the insights acquired throughout the Analyze and Construct phases. Additionally, an executive summary is presented to summarize key findings and recommendations, ensuring clear communication of the project's outcomes and how they align with the business objectives. This phase ensures that stakeholders are informed and can make decisions based on the insights and models developed during the project.

7. **Compute descriptive statistics:** Analyze

Why did you select this stage for this task?

This is done in the Analyze phase where statistical analysis is performed as part of hypothesis testing and overall data exploration (EDA) to understand the data. Additionally, it's also done in this phase to establish the key insights that guide the selection of the right algorithm. This is crucial for model training, ensuring that the selected approach aligns with the patterns and trends identified in the data, ultimately leading to more accurate and reliable model outcomes.

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

Why did you select these stages for this task?

This is a task that is done in the Execute phase of the PACE strategy. In this phase, the visualizations, along with the inferred insights and executive summary, are shared with stakeholders. This provides them with an overview of the findings, allowing for feedback, suggestions, or additional inputs to be gathered. These inputs are then incorporated into the final deliverables, ensuring that the results meet stakeholders' expectations and contribute to informed decision-making.

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

Why did you select this stage for this task?

This is one of the tasks in the starting phase of the PACE workflow, as a project proposal establishes the scope, objectives, requirements, goals, deliverables, milestones, tasks, etc. This document serves as a foundational reference to stay on track and provides an overview of the project for anyone who joins the team. Additionally, if any changes take place, the proposal can be updated and compared to the original version, ensuring alignment and clarity throughout the project's lifecycle.

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

Why did you select this stage for this task?

This is a model-building task and it belongs to the construct phase clearly. In this phase, the model is first initialized with some parameter values, and then the desired algorithm, along with the training data, is provided. The model follows the algorithm's process to infer patterns and relationships in the training data, allowing it to begin learning and making predictions. Additionally, the analyzing stage will examine the model in detail to ensure it will meet the needs of the task, validating its effectiveness before it is fully deployed.

11. **Compile summary information about the data:** Analyze

Why did you select this stage for this task?

Inspecting a dataset to compile information takes place during the analysis phase of the PACE framework. Calculating summary statistics is a key part of this phase, specifically during the Exploratory Data Analysis (EDA) process. This step involves quantifying and summarizing the data's characteristics, helping to uncover insights and patterns that will guide further analysis and model development.

12. **Build machine learning model:** Construct

Why did you select this stage for this task?

The building of a data model takes place in the construct phase, as this stage involves developing the model based on the decisions made from the insights inferred during the analyze phase. In this phase, the selected algorithm is used to guide the model's development, and the model is trained on the data to learn patterns and relationships. The process ensures that the model aligns with the objectives and insights derived from the earlier stages of the project.