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?

Effectively navigating this project requires engaging with two primary groups: firstly, the internal Automatidata team (Udo Bankole, Deshawn Washington, Luana Rodriquez, Uli King), who possess technical expertise and value concise updates; and secondly, the external NYC TLC client contacts (Juliana Soto, Titus Nelson), who operate in non-technical management roles and therefore require information presented with clarity, accessible language, and explanations tailored to their operational perspective.

* 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 specific objective for this Course 1 assignment is to develop a detailed and actionable project proposal for the Automatidata team, structuring the approach using the PACE framework, defining project milestones, identifying necessary early-phase tasks like data acquisition, cleaning, and exploratory analysis, clarifying stakeholder roles and communication plans, and outlining preliminary timelines and risks, ultimately creating a foundational document to align the team for the NYC TLC fare prediction project. Looking ahead, the anticipated impact of successfully delivering the overall fare prediction model for the NYC TLC is multifaceted, aiming to significantly enhance customer experience through reliable pre-ride estimates (increasing fare transparency), equipping the TLC with data-driven insights into the complex factors determining fares to inform policy or fare structure reviews, and potentially supporting improved operational efficiencies and more effective regulatory oversight.

* What questions need to be asked or answered?

Before starting the core analysis for the NYC TLC project, several key questions need clarification: What is the exact target variable (e.g., fare\_amount vs total\_amount) and what specific performance metrics will define the success of the final prediction model according to the TLC? What is the precise scope of the provided dataset (timeframe, known limitations), and crucially, how do the PULocationID and DOLocationID codes map to actual geographic zones (is a lookup table available)? We also need to confirm which predictor variables will realistically be available before a ride starts, understand the requirements for the executive-level visualizations requested by Titus Nelson, and ascertain the key priorities or concerns of both TLC stakeholders (Juliana Soto and Titus Nelson) regarding the project's outcomes.

* What resources are required to complete this project?

Successfully completing the NYC TLC fare prediction project for Automatidata requires several key resources: primarily, access to the NYC TLC dataset (2017\_Yellow\_Taxi\_Trip\_Data.csv) and potentially a Taxi Zone lookup table for geographic analysis. Essential software and tools include a Python environment with core data science libraries (Pandas, NumPy, Matplotlib, Seaborn, Scikit-learn) likely run via Jupyter Notebooks, plus Tableau for specific visualizations and standard office software for documentation. Crucially, personnel resources involve the expertise and collaboration of the Automatidata data team (Director, Manager, Senior Analysts, Project Manager, yourself) and scheduled access to the NYC TLC stakeholders (Juliana Soto, Titus Nelson) for clarifications and feedback. Lastly, adequate time allocated for each project phase and potentially sufficient computing power for data processing and model training are necessary.

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

Throughout the complete lifecycle of the NYC TLC fare prediction project, key deliverables will include initial planning documents like the PACE Strategy Document and the Project Proposal (focus of Course 1), followed by outputs from the analysis phase such as a cleaned dataset and a comprehensive Exploratory Data Analysis (EDA) report (likely a documented Python notebook with numerous visualizations, the focus of Course 3). Subsequently, the model building phase will produce the trained fare prediction regression model itself, along with documentation on feature engineering and training. Finally, the execution phase requires a model evaluation report detailing performance metrics, and communication deliverables such as a final Executive Summary and potentially a presentation incorporating key visuals tailored for both technical (Automatidata) and non-technical (TLC) 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](https://www.coursera.org/learn/foundations-of-data-science/supplement/4OtHr/the-pace-stages) and [Communicate objectives with a project proposal](https://www.coursera.org/learn/foundations-of-data-science/supplement/79Ysh/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?

Model evaluation fits into the Execute stage as this assessment occurs after the model development (Construct phase) is finished. It involves applying the model to unseen data (like a test set) to rigorously measure its performance against the project's predefined success criteria and business objectives, determining its effectiveness and readiness.

1. Conduct hypothesis testing: Analyze and Construct

Why did you select these stages for this task?

Hypothesis testing typically spans two stages: during the Analyze stage, the need for a specific statistical test is identified based on data exploration and questions arise, leading to the formulation of hypotheses; subsequently, the Construct stage involves the practical application and computation of that statistical test on the relevant data.

1. Begin exploring the data: Analyze

Why did you select this stage for this task?

The initial exploration of the data belongs squarely in the Analyze stage. This phase is dedicated to diving into the dataset using descriptive statistics, visualizations, and other EDA techniques to thoroughly understand its structure, identify patterns, check quality, and extract preliminary insights.

1. Data exploration and cleaning: Plan and Analyze

Why did you select these stages for this task?

This task connects two stages: preliminary Planning occurs when outlining the overall approach and deciding which specific exploration techniques and cleaning methods (e.g., for handling outliers or missing values) will be necessary based on project goals and initial data assessment. The subsequent hands-on execution of these cleaning processes and detailed exploratory analysis then takes place within the Analyze stage.

1. Establish structure for project workflow (PACE): Plan

Why did you select this stage for this task?

Defining the project's overall structure using the PACE workflow occurs in the Plan stage, as this initial phase involves creating the foundational documents, like the PACE strategy itself, which outline the sequence and approach for all subsequent tasks and helps ensure alignment

1. Communicate final insights with stakeholders: Execute

Why did you select this stage for this task?

The communication of finalized insights and project outcomes to stakeholders is a primary activity of the Execute stage, typically occurring after model evaluation and interpretation are complete, representing the formal delivery of the project's findings.

1. Compute descriptive statistics: Analyze

Why did you select this stage for this task?

Calculating and investigating descriptive statistics (such as mean, median, min, max, standard deviation) is a core practice within the Analyze stage, essential for quantitatively understanding the characteristics and distributions of the data variables.

1. Visualization building: Analyze and Construct

Why did you select these stages for this task?

Visualization building connects two phases: initial visual exploration to understand data and determine which charts are needed occurs during Analyze, while the technical creation and coding of these planned visualizations using software tools happens during the Construct stage.

1. Write a project proposal: Plan

Why did you select this stage for this task?

Crafting the project proposal falls definitively within the Plan stage, as this document serves as the crucial initial blueprint that formally defines the project's objectives, scope, methodology, deliverables, timelines, and stakeholders.

1. Build a regression model: Analyze and Construct

Why did you select this stage for this task?

Developing the regression model spans these stages: the Analyze stage includes the in-depth data exploration required to select appropriate features and determine if a regression model meets the task needs, whereas the Construct stage involves the actual coding, training, tuning, and iterative building of the model itself.

1. Compile summary information about the data: Analyze

Why did you select this stage for this task?

Compiling summary information falls into the Analyze stage because this involves actively inspecting the dataset, calculating descriptive statistics, and synthesizing key details about the data's properties, which are core exploratory analysis activities.

1. Build machine learning model: Construct

Why did you select this stage for this task?

The actual development and implementation of a machine learning model, including writing the code, training the algorithm on data, and tuning its parameters, is the central activity of the Construct stage in the PACE workflow.