

# Automatidata PACE Strategy Document I

## High-level Planning

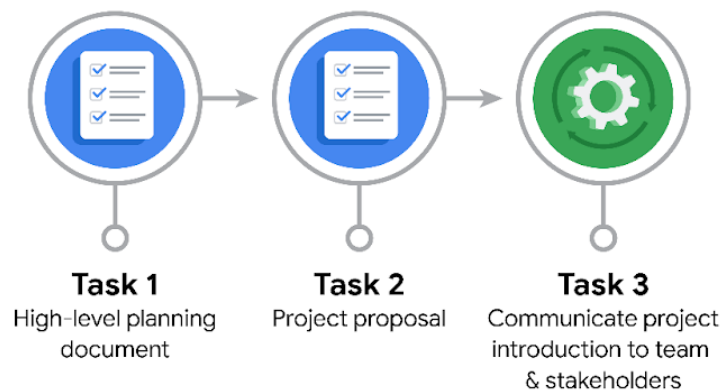
### Introduction

PACE stands for Plan, Analyze, Construct and Execute. It is a framework that illustrates the foundation and structure for data analysis projects and each letter represents an actionable stage in a project. The stage “Plan” involves the definition of the project scope, the research of business data and the workflow development. The stage “Analyze” involves data scrubbing, data conversion and database formatting. The stage “Construct” involves building models and machine learning algorithms and selecting a modeling approach. The stage “Execute” involves the presentation of results to decision-makers, stakeholders and others in order to receive feedback. This framework is built upon an iterative cycle where each stage may reveal new insights, requiring the return to earlier stages. A PACE strategy document is used to record decisions and reflections at different stages of the data analytical process. It typically includes the definitions of roles and actions to ensure clarity and accountability.



### Purpose

Automatidata is consulting for the New York City Taxi and Limousine Commission (TLC). New York City TLC is an agency responsible for licensing and regulating New York City's taxi cabs and for-hire vehicles. The agency has partnered with Automatidata to develop a regression model that helps estimate taxi fares before the ride, based on data that TLC has gathered. For this stage of the project, we identify three main tasks that are presented in the following visual.





## Considerations



### PACE: Plan Stage

- Who is the target audience for this project?

The primary audience for this project is the New York City Taxi and Limousine Commission.

- What is the objective or problem we aim to address?

The problem we are trying to solve is the accurate estimation of taxi fares based on relevant variables.

- What are the key questions that must be addressed?

Key considerations include:

- What is the quality and condition of the available dataset?
- Which variables will have the most significant impact on fare predictions?
- Are there identifiable trends within the data that can guide our analysis?
- How can we mitigate potential bias in the data and analysis process?

- What resources are essential for completing this project?

We will need Python scripts and notebooks, the project dataset and ongoing stakeholder input.

- What deliverables must be produced during the project lifecycle?

Deliverables include:

- Cleaned and prepared dataset for exploratory data analysis (EDA)
- Data visualizations to highlight insights
- Statistical model
- Regression model

- What is the timeline for the project?


The project is expected to be completed within 7 weeks, with specific milestones.

- Who are the key stakeholders involved in decision-making?

The stakeholders include representatives from New York City TLC and members of the data team.

- What risks might impact project completion?

Potential risks include data quality issues and model performance falling short of expectations.

- 
- How will the project success be measured?

Success will be measured by:

- Accuracy of the fare predictions
- Quality of the insights generated from the data
- Stakeholder satisfaction with the final results

## Tasks

The following group of tasks needs to be completed within the project. We identify which stage of the PACE workflow each task would best fit.

### 1. Draft the project proposal: **Plan**

Creating a detailed project proposal is foundational as it sets the direction for the entire project.

### 2. Design project workflow structure: **Plan**

Establishing a project workflow early with an initial PACE document ensures that processes are well-defined, roles are clear and timelines are realistic.

### 3. Summarize key data characteristics: **Analyze**

Summarizing the data is an early step of the analysis phase to gather initial understanding of the dataset's composition, types and sources.

### 4. Initiate data exploration: **Analyze**

The analysis phase provides a deeper understanding of the dataset and the information within it. We want to assess its potential, identify patterns and recognize any obvious challenges in the dataset.

### 5. Clean and analyze the data: **Plan** and **Analyze**

Planning takes place when we make choices about the methods needed. Data cleaning and deeper exploration happen concurrently, where the raw data is refined for analysis and patterns are investigated.

### 6. Create data visualizations: **Analyze** and **Construct**

Visualization begins with data assessment and is created during the construction stage, summarizing key insights in a more understandable format for both the data team and stakeholders.



**7. Calculate descriptive statistics: Analyze**

Descriptive statistics help quantify and summarize key features of the data, providing a basis for further analysis.

**8. Perform hypothesis testing: Analyze and Construct**

A statistical test is chosen and prepared during the analysis phase, but the test is carried out during the construction stage.

**9. Develop regression model: Analyze and Construct**

A detailed model examination is done during the analysis phase. Building the regression model is a core task of the construction stage.

**10. Assess model performance: Execute**

Evaluating the model after its construction involves checking its accuracy, fit and validity, which is crucial to ensure it meets project objectives.

**11. Deliver final insights to stakeholders: Execute**

Communication is important at various points throughout a project, but presenting the findings and their implications to decision-makers and stakeholders for approval and action is part of the execution phase.