

# Automatidata PACE Strategy Document II

## *Preliminary Data Summary*

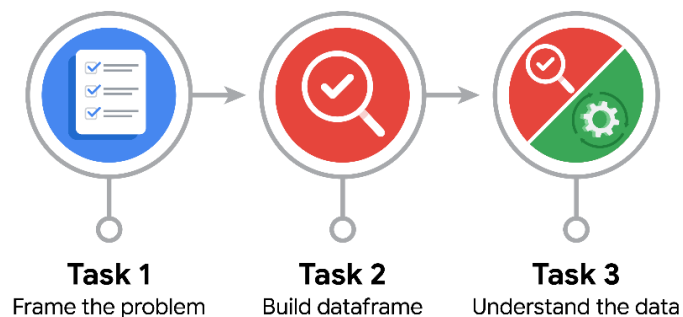
### Introduction

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

- How can we best prepare to understand and organize the provided data?

The first step is to explore the dataset and consider reviewing the Data Dictionary. By analyzing the taxi cab data fields, we can better understand the significance of each variable. However, the main objective is to load the data into Python, examine it closely and share initial observations with the Data Analysis Manager. The next step would be to deepen our understanding and check for any irregularities in the data.

- What follow-along and self-review materials can assist in completing this task?

Helpful resources include the Data Dictionary, which explains each variable in the dataset, and the fact sheet, which provides background information on the data source. Additionally, reviewing documentation on descriptive statistics and regression analysis will support a better understanding of how to summarize and interpret the dataset.

- What steps might a proactive learner take before starting to code?


A proactive learner may begin by thoroughly reviewing the data and understanding the structure and meaning of each variable. They could also explore similar datasets or past analyses for guidance. Before coding, it is beneficial to identify key variables, consider potential relationships between them, and take note of any patterns or anomalies that may need special attention.



### PACE: Analyze Stage

- Is the information provided enough to reach the goal based on our intuition and variable analysis?

The available data seems sufficient to build a predictive model for taxi fares, especially with key variables such as the total amount and the trip distance. However, further analysis is necessary to ensure the accuracy of the data and to identify any outliers or anomalies that could skew the results.

- 
- How would we create a summary dataframe and determine the minimum and maximum values within the dataset?

To build a summary dataframe, the basic descriptive statistics functions in Python (e.g. *describe()* in *pandas*) would be used. This would provide essential metrics like mean, median and standard deviation, as well as the minimum and maximum values for each variable. Filtering the data to check for specific ranges and inspecting the distribution of values can help in identifying anomalies.

- Do any average values appear out of the ordinary? How would we describe the interval data?

In the preliminary inspection, some average values, particularly in the variable for the total amount, appeared unusual due to discrepancies in charges for short-distance trips. This indicates the presence of outliers. The interval data, which is continuous, can be described by its range and central tendency (e.g. mean and median), with attention to any irregularities that deviate from the expected patterns.



### **PACE: Construct Stage**

**Note:** This stage is not relevant to the current workflow.



### **PACE: Execute Stage**

- Based on your current understanding of the data, what initial recommendations would we provide our manager for further investigation before diving into exploratory data analysis (EDA)?

It would be advisable to investigate the outlier trips with unusually high fares for short distances, as these could skew the results of the predictive model. A closer look at the distribution of variables is also recommended to ensure the data's integrity before further analysis.

- Which data points seem to contain irregularities?

The variable for the total amount includes several trips with disproportionately high fares for short distances, which suggests the presence of outliers. These data points need to be examined closely to determine whether they are valid or the result of data entry errors or other factors.



- What additional types of data could enhance this dataset?

Additional data such as weather conditions or special events could provide more context for understanding taxi ride fares. Including driver or vehicle information might also help in identifying patterns that affect pricing, such as peak hours and demand surges.