**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?

My audience for this project are the team member of the New York City Taxi and Limousine Commission (TLC) ~~and the team members of Automatidata.~~

* 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?

We want to develop a regression model that estimates taxi fares before a customer takes a ride with one of the taxis licensed and regulated by TLC.

By being able to provide fare estimations to their users, TLC will make the experience of riding with one of their licensed taxis more attractive to their customers, and in this way might lead to an increase in rides.

* What questions need to be asked or answered?

What data related to taxi rides has TLC collected and can share with Automatidata?

Is that data available without any personally identifiable information, or does it need to be anonymized or aggregated before it can be used for the project?

Will the data provided by TLC be provided in a cleaned state?

Who will be the point of contact at TLC in case we might have questions about the data?

Will additional data from other sources need to be collected in order to create an accurate model?

How accurate should the model be in order to consider the project a success?

What will be the level of involvement (RACI) of the Automatidata team members for each stage of the project?

What is the deadline for completing this project?

What variables will be the most useful? Are there trends within the data that can provide insight? What

steps can I take to reduce the impact of bias?

* What resources are required to complete this project?

Data provided by TLC

Software and infrastructure for managing the data and training machine learning models

Time resources of the members of the Automatidata team

Python notebook, and input from stakeholders

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

Cleaned data ready to be used by machine learning models

A trained machine learning models which estimates ride fares based on a set of input parameters

Description of the results/performance of the machine learning model, including visualizations

An executive summary

The deliverables include a dataset scrubbed for exploratory data analysis,

visualizations, statistical model, regression analysis and/or machine learning model.

## 

## **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:** Construct

Why did you select this stage for this task?

The work on the machine model, including its testing and evaluation, happens during the Construct stage.

After the model has been constructed, data is run through to evaluate whether it meets the project’s expectations and goals.

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

Why did you select these stages for this task?

Hypothesis testing requires deciding first what data is relevant for solving the problem at hand. Then, a hypothesis must be made based on the data retained after having conducted the exploratory data analysis, which happens during the Analyze stage. Finally, the hypothesis is tested during the Construct stage, when a model is developed and tested.

The outcome of Construct stage will determine whether the hypothesis is accepted. If the hypothesis is refuted, then we might have to return to the Analyze stage to collect more data before repeating the Construct stage.

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

Why did you select this stage for this task?

The interaction with the data starts in the Analyze stage.

1. **Data exploration and cleaning:** Analyze **and Select PACE stage**

Why did you select these stages for this task?

The data exploration and cleaning is carried out during the Analyze stage. The cleaned data is a deliverable of that stage.

Planning takes place when you first make choices about the methods needed. The cleaning process then takes place in the analyzing stage.

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

Why did you select this stage for this task?

Everything related to project planning and organizational aspects happens during the Plan stage of the project.

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

Why did you select this stage for this task?

The communication of results and potential recommendations is done during the last stage of a project, which is the Execute stage.

1. **Compute descriptive statistics:** Analyze

Why did you select this stage for this task?

Descriptive statistics are typically computed while the exploratory data analysis is carried out, which occurs during the Analyze stage.

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

Why did you select these stages for this task?

Visualizations are created while analyzing the data (during the Analyze stage) and while developing and testing machine learning algorithms during the Construct stage.

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

Why did you select this stage for this task?

The project proposal must be prepared at the beginning of the project, during the Plan stage.

1. **Build a regression model:** Construct **and Select PACE stage**

Why did you select this stage for this task?

The modeling work is carried out during the Construct stage.

During the analyzing stage, the model is examined in detail to be sure it will meet the needs of the task. The building of the regression model will take place in the construction phase.

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

Why did you select this stage for this task?

The summary information can be helpful to communicate to the data team once the EDA has been performed and the descriptive statistics have been computed. These happen during the Analyze stage.

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

Why did you select this stage for this task?

The work on machine learning models is carried out during the Construct stage, for which the main deliverable is the machine learning model. However, for this project a regression model will be developed, so developing another machine learning model is not necessary.