**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?

**Udo Bankole** (Director of Data Analysis): Needs detailed insights and final presentation talking points.

**Deshawn Washington** (Data Analysis Manager): Oversees project tasks and progress.

**Luana Rodriquez** (Senior Data Analyst): Involved in data inspection, analysis, and model testing.

**Uli King** (Senior Project Manager): Manages project milestones and communication with stakeholders.

**Juliana Soto** (Finance and Administration Department Head): Requires understandable summaries and explanations of the project's financial and operational impacts.

**Titus Nelson** (Operations Manager): Interested in operational insights and practical implications of the project outcomes.

**Executives and Decision-Makers** at the NYC TLC: They need high-level insights and visualizations to make informed decisions based on the project's findings and recommendations.

* 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?

**Predict Taxi Fares**: Develop a regression model to accurately estimate taxi fares before a ride based on factors such as distance, time of day, and other relevant variables.

**Improve Operational Efficiency**: Utilize the predictive model to streamline operations for the NYC TLC, optimizing fare estimations and reducing discrepancies.

**Provide Strategic Insights**: Analyze the TLC dataset to identify trends and patterns that can inform strategic decision-making and policy adjustments.

**Enhanced Customer Experience**: By providing accurate fare estimates, customers will have greater transparency and trust in the pricing of taxi services.

**Operational Cost Savings**: Accurate fare predictions can help reduce operational inefficiencies, leading to cost savings for both the TLC and taxi operators.

**Data-Driven Decision Making**: The insights gained from exploratory data analysis and the regression model can guide the TLC in making informed decisions regarding fare structures, operational policies, and resource allocation.

**Increased Stakeholder Satisfaction**: Effective communication of project findings and recommendations will ensure that both technical and non-technical stakeholders are aligned, leading to higher satisfaction and support for the project's outcomes.

**Foundation for Future Projects**: The methodologies and frameworks established in this project can serve as a blueprint for future data-driven initiatives within the TLC and other similar organizations.

By addressing these objectives and anticipating these impacts, the project aims to not only solve the immediate problem of fare estimation but also contribute to the larger strategic goals of the NYC TLC.

* What questions need to be asked or answered?

What is the structure and format of the TLC dataset?

Are there any missing or incomplete data points?

How can we clean and preprocess the data to ensure accuracy and reliability?

Which variables (e.g., distance, time of day, location) are available in the dataset?

How do these variables correlate with taxi fares?

Are there any additional variables that should be considered for the regression model?

What type of regression model will best suit the prediction of taxi fares?

How will we split the data for training and testing the model?

What metrics will we use to evaluate the model's performance?

What trends and patterns can be identified in the data?

Are there any outliers or anomalies that need to be addressed?

How can visualizations aid in understanding the data distribution?

What information do the technical team members need to proceed with the analysis?

How can we present the findings and recommendations in a way that is understandable to non-technical stakeholders?

What are the key insights and talking points to highlight in the final presentation to TLC executives?

How will the predictive model affect the operational efficiency of the TLC?

What are the potential cost savings or revenue implications of accurate fare predictions?

What are the technical requirements for deploying the regression model in a real-time environment?

How will we monitor and maintain the model's performance over time?

Are there any privacy or security concerns related to the use of the TLC dataset?

How can we ensure that the model's predictions are fair and unbiased?

How will we gather feedback from stakeholders on the model's performance and usefulness?

What processes will we put in place for continuous improvement and iteration of the model?

By addressing these questions, we can ensure a comprehensive approach to the project, covering all necessary aspects from data preparation to stakeholder communication and strategic impact.

* What resources are required to complete this project?

Data and Technical Resources

TLC Dataset:

Access to the complete dataset from the New York City Taxi and Limousine Commission.

Data includes variables such as trip distance, time of day, pickup and drop-off locations, fare amounts, and any additional relevant data points.

Software and Tools:

Python for data analysis, model building, and testing.

Jupyter Notebooks or other IDEs (e.g., PyCharm, VSCode) for coding and documentation.

Pandas, NumPy, Scikit-Learn, and other relevant Python libraries for data manipulation, analysis, and machine learning.

Matplotlib and Seaborn for data visualization.

SQL for data querying if the dataset is stored in a relational database.

Git and GitHub for version control and collaboration.

Computing Resources:

High-performance computing environment or cloud-based resources (e.g., Google Colab, AWS, Azure) for handling large datasets and performing intensive computations.

Human Resources

Project Team Members:

Data Analysts (e.g., Luana Rodriquez): Responsible for data cleaning, exploratory data analysis, and visualization.

Data Scientists (e.g., Udo Bankole): Focused on model development, testing, and evaluation.

Project Managers (e.g., Uli King): Overseeing project milestones, timelines, and communication with stakeholders.

Data Analysis Manager (e.g., Deshawn Washington): Providing overall direction and ensuring alignment with project goals.

Stakeholder Communication:

Regular meetings and updates with TLC team members (Juliana Soto, Titus Nelson) to gather requirements, provide progress updates, and present findings.

Documentation and Templates

PACE Strategy Document:

Template to guide the planning, analysis, construction, and execution stages of the project.

Project Proposal Template:

Template for organizing and communicating project tasks, milestones, and stakeholder roles.

Training and Knowledge

Training Materials:

Access to relevant tutorials, courses, and documentation on regression modeling, data analysis, and Python programming.

Domain Knowledge:

Understanding of the taxi and transportation industry to better interpret the data and provide actionable insights.

Infrastructure

Data Storage:

Secure storage solutions for the TLC dataset, ensuring data privacy and compliance with relevant regulations.

Collaboration Tools:

Communication platforms (e.g., Slack, Microsoft Teams) for team collaboration and coordination.

Project management tools (e.g., Asana, Trello) to track tasks, milestones, and deadlines.

By ensuring these resources are in place, the project team can effectively execute the tasks required to develop the predictive model and deliver valuable insights to the NYC TLC.

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

### Project Deliverables

#### Initial Planning

1. \*\*Project Proposal\*\*:

- A comprehensive document outlining the project’s objectives, tasks, milestones, and stakeholder roles.

- Includes a detailed timeline and resource allocation plan.

2. \*\*PACE Strategy Document\*\*:

- A structured plan using the PACE framework to organize project tasks and stages (Plan, Analyze, Construct, Execute).

#### Data Preparation and Analysis

3. \*\*Data Quality Report\*\*:

- A report detailing the initial assessment of the TLC dataset, including data cleaning, preprocessing steps, and any data quality issues identified.

4. \*\*Exploratory Data Analysis (EDA) Report\*\*:

- Visualizations and statistical summaries to uncover trends, patterns, and insights within the data.

- Identification of key variables for the regression model.

#### Model Development and Testing

5. \*\*Regression Model\*\*:

- Development of a regression model to predict taxi fares based on the identified variables.

- Includes model selection, training, validation, and testing processes.

6. \*\*Model Evaluation Report\*\*:

- Detailed evaluation of the model’s performance using appropriate metrics (e.g., RMSE, R-squared).

- Discussion of model accuracy, reliability, and any potential limitations.

7. \*\*A/B Testing Results\*\*:

- Analysis of the relationship between key variables and the impact of different scenarios using A/B testing.

#### Implementation and Deployment

8. \*\*Model Deployment Plan\*\*:

- Strategy for deploying the regression model in a real-time environment, including technical requirements and monitoring processes.

#### Communication and Reporting

9. \*\*Interim Progress Reports\*\*:

- Regular updates on project progress, challenges, and milestones for both internal team members and external stakeholders.

10. \*\*Stakeholder Presentation\*\*:

- A comprehensive presentation of the project’s findings, model performance, and strategic recommendations tailored for TLC executives and decision-makers.

- Visualizations and summaries to communicate insights clearly and effectively.

#### Final Deliverables

11. \*\*Final Project Report\*\*:

- A detailed document summarizing the entire project lifecycle, from initial planning to model deployment and evaluation.

- Includes methodologies, findings, insights, and recommendations for future actions.

12. \*\*Project Documentation\*\*:

- Complete documentation of all code, processes, and methodologies used throughout the project.

- Ensures reproducibility and serves as a reference for future projects.

#### Optional (Based on Project Needs)

13. \*\*Interactive Dashboard\*\*:

- A user-friendly dashboard to visualize key insights and allow stakeholders to interact with the data and model predictions.

14. \*\*Training and User Guide\*\*:

- Documentation and training materials for TLC staff to understand and utilize the predictive model and associated tools.

By systematically delivering these outputs, the project team can ensure comprehensive coverage of all necessary tasks and provide valuable deliverables to meet the NYC TLC's needs.

## 

## **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 task of evaluating the model fits best under the "Analyze" stage of the PACE workflow because this stage focuses on assessing and interpreting data and models to derive insights and ensure accuracy. Evaluating the model involves measuring its performance using various metrics, validating its accuracy, reliability, and consistency, and determining whether it meets the project's requirements. This analysis is crucial for understanding the effectiveness of the model and making any necessary adjustments to improve its predictive capabilities. In essence, this task is about analyzing the outcomes of the model to ensure it aligns with the project goals and can be reliably used for making predictions.

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

Why did you select these stages for this task?

Plan:

Conducting hypothesis testing begins in the "Plan" stage because it involves defining the research question, formulating hypotheses (null and alternative), and determining the appropriate statistical methods and tests to use. This stage sets the groundwork for the analysis by identifying the objectives and the data needed to test the hypotheses. Proper planning ensures that the hypothesis testing is aligned with the project's goals and that the approach is methodologically sound.

Analyze:

The actual execution of hypothesis testing occurs in the "Analyze" stage, where data is examined to test the formulated hypotheses. This involves applying statistical tests, interpreting the results, and determining whether to reject or fail to reject the null hypothesis. The analysis provides insights into the relationships and patterns within the data, helping to draw conclusions and inform decision-making based on the statistical evidence gathered.

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

Why did you select this stage for this task?

Exploring the data is a key part of the "Analyze" stage because it involves examining the dataset to understand its structure, patterns, and initial insights. This task includes activities like visualizing data distributions, identifying missing values, and understanding the relationships between variables. Data exploration sets the foundation for deeper analysis by providing an initial understanding of the data’s characteristics.

1. **Data exploration and cleaning: Analyze** **and Construct**

Why did you select these stages for this task?

Analyze:

Data exploration is a part of the "Analyze" stage because it involves investigating the dataset to identify trends, patterns, anomalies, and relationships among variables. This stage helps in gaining a deeper understanding of the data, which is crucial for further analysis.

Construct:

Data cleaning fits into the "Construct" stage as it involves transforming the data by handling missing values, correcting errors, and ensuring data consistency and quality. This stage prepares the data for accurate and reliable analysis and modeling.

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

Why did you select this stage for this task?

Establishing the structure for the project workflow is a "Plan" stage activity because it involves defining the project's roadmap, setting objectives, milestones, and timelines, and assigning roles and responsibilities. This stage ensures that the project is well-organized and that all team members are aligned on the project's goals and processes.

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

Why did you select this stage for this task?

Communicating final insights with stakeholders falls under the "Execute" stage as it involves delivering the final results and findings to the relevant audience. This task includes creating reports, presentations, and visualizations to effectively convey the insights derived from the analysis. The execution stage is where the results are put into action and shared with decision-makers.

1. **Compute descriptive statistics: Analyze**

Why did you select this stage for this task?

Computing descriptive statistics is part of the "Analyze" stage because it involves summarizing the main features of the dataset through measures such as mean, median, standard deviation, and other statistical summaries. These statistics provide a quick overview of the data and help in understanding its basic properties before further analysis.

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

Why did you select these stages for this task?

Analyze:

Visualization building initially falls under the "Analyze" stage because it involves creating visual representations of the data to uncover patterns, trends, and relationships that are not easily seen in raw data.

Construct:

It also fits in the "Construct" stage as developing effective visualizations requires constructing visual tools and dashboards that can be used to communicate findings clearly and effectively.

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

Why did you select this stage for this task?

Writing a project proposal is a "Plan" stage activity because it involves outlining the project's objectives, scope, methodology, and timeline. The proposal sets the foundation for the entire project by detailing how it will be executed and what resources will be needed.

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

Why did you select this stage for this task?

Construct:

Building a regression model involves developing the actual predictive model, which fits into the "Construct" stage as it includes the technical work of coding, training, and refining the model.

Analyze:

It also fits in the "Analyze" stage because once the model is built, it must be evaluated and tested to ensure it accurately predicts outcomes based on the data. This involves interpreting the model’s performance and making necessary adjustments.

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

Why did you select this stage for this task?

Compiling summary information about the data is an "Analyze" stage task because it involves summarizing the key characteristics and insights from the data exploration and analysis. This summary helps in understanding the data better and is crucial for further steps in the analysis and modeling process.

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

Why did you select this stage for this task?

Building a machine learning model is part of the "Construct" stage because it involves the technical work of developing, training, and fine-tuning a model to learn from the data. This stage focuses on creating the model that will be used for predictions and analysis.