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.

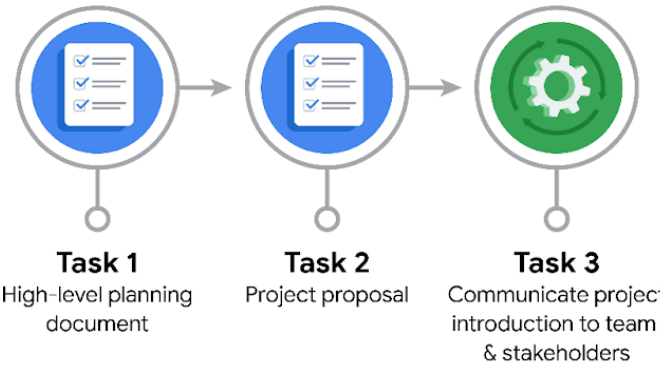
Relevant Interview Questions

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?

Automatidata's data team:
- Deshawn Washington (Data Analysis Manager)
- Luana Rodriquez (Senior Data Analyst)
- Udo Bankole (Director of Data Analysis)
- Uli King (Senior Project Manager)

New York City TLC team members:
- Juliana Soto (Finance and Administration Department Head)
- Titus Nelson (Operations Manager)

- 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?

What Automatidata is trying to solve or accomplish:

- **Problem:** The New York City Taxi and Limousine Commission (TLC) wants to provide a better experience for riders by offering fare estimates before trips, but they lack a tool to do this.

- **Solution:** Automatidata will develop a predictive model, specifically a regression model, to accurately estimate taxi fares. This model will then be integrated into a user-friendly application (app) for TLC riders.
- **Accomplishment:** Deliver a functional app that provides reliable fare estimations, empowering riders with greater transparency and control.

Anticipated impact on the larger needs of the TLC:

- **Enhanced Rider Experience:** The app will significantly improve the rider experience by reducing uncertainty about fares, leading to increased satisfaction and trust.
- **Increased Transparency and Trust:** By providing clear fare estimates, the TLC will demonstrate its commitment to transparency, fostering trust between riders and the agency.
- **Improved Operational Efficiency:** The app could potentially reduce the number of fare-related disputes and inquiries, freeing up TLC resources.
- **Data-Driven Insights:** The data collected and analyzed for the model can provide valuable insights into rider behavior and fare dynamics, which can inform future TLC policies and Initiatives.
- **Modernization and Innovation:** Implementing a fare estimation app positions the TLC as a modern and innovative agency that leverages technology to serve its riders better.
- **Potential for Increased Ridership:** The improved rider experience and transparency could encourage more people to use taxis, potentially increasing ridership.

- What questions need to be asked or answered?

1. Project Goals & Scope:

A. What are the specific requirements for the fare estimation app from the TLC's perspective?
Does the app need to provide a precise fare, or a fare range?
Should the app account for surge pricing or traffic conditions?
B. What are the key variables that influence taxi fares (distance, time, location, etc.)?
Are there time-of-day or day-of-week surcharges?
Does the type of vehicle (e.g., standard taxi, luxury vehicle) affect the fare?
Does the pickup or drop-off location (e.g. zones, airport) affect the fare?
C. What is the timeline for the project?

2. Data & Model:

A. What specific data fields are available in the TLC dataset, and what is the quality of the data (missing values, outliers, inconsistencies)?
- Duplicates, missing values, bias,...
B. How accurate does the model need to be to meet TLC's requirements?
What is the margin error?
C. Which regression model is most appropriate for this data?

3. Stakeholder & Communication:

A. What are the TLC's expectations for the project?
Does the TLC expect regular progress reports, and in what format?
Does the TLC expect training or documentation for the app?
B. How will we keep all stakeholders informed of progress?
Will we hold weekly progress meetings with the TLC team?
Will we use a project management tool to track tasks and progress?
Who will be the primary point of contact for communications?
C. What are the main priorities of each stakeholder?

- What resources are required to complete this project?

1. Data sources:
A. TLC Taxi Trip Data
- Dataset containing historical trip information (pickup/drop-off locations, times, fares, distances,...)
- Access to any data documentation describing the data.

B. External data
- Traffic data (historical and real time)
- Weather conditions
Does rainfall correlate with increased customer volume?
- Geographic data, if fares vary by location.

2. Technical Resources:
A. Hardware:
- Computers with sufficient processing power and memory for data analysis and model training.
- Data storage infrastructure (local or cloud-based).

B. Software:
Python:
- Pandas (for data manipulation).
- Scikit-learn (for machine learning models).
- Matplotlib and Seaborn (for data visualization).
- Numpy (for numerical computing)

Data Storage/Database:
- Cloud based storage like Google Cloud Storage.
- Databases such as PostgreSQL or MySQL.

Integrated Development Environment (IDE):
Jupyter Notebooks or VS Code (for coding).

C. Cloud Computing (Optional):
Cloud platforms like Google Cloud Platform (GCP) for scalable data processing and model training.

3. Human Resources:
A. Data Professionals (Analysts/Scientists):

Individuals with expertise in data cleaning, EDA, regression modeling, and model evaluation.

B. Project Manager:

To oversee the project, manage timelines, and ensure effective communication

C. Communication Specialists (Optional):

To aid in the creation of reports and presentations for non-technical stakeholders.

4. Communication and Collaboration Resources:

A. Communication Tools:

Email, instant messaging (e.g., Slack, Microsoft Teams).
Video conferencing (e.g., Zoom, Google Meet).

B. Project Management Tools:

Tools for task tracking, progress monitoring, and document sharing

C. Document Sharing and Collaboration:

Google Drive, or similar platforms.

5. Time and Budget:

Time:

Adequate time for each phase of the project:
(data collection, analysis, model building, communication).

Budget:

Funds for software licenses, cloud computing resources, and potential external data purchases.

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

1. Planning & Setup Phase:

A. Project Proposal:

A comprehensive document outlining the project's goals, scope, timeline, milestones, resources, and communication plan.

B. *PACE Strategy Document:*

A document detailing the project's planning, analysis, construction, and execution phases.

2. Data Analysis & Exploration Phase:

Data Exploration Report (EDA):

A report summarizing the initial analysis of the TLC dataset, including data quality assessments, descriptive statistics, and visualizations.

Cleaned Dataset:

A processed and cleaned version of the TLC dataset, ready for model building.

Feature Engineering Documentation:

A document detailing the creation of new features and the selection of relevant variables for the model.

3. Model Building & Evaluation Phase:

Trained Regression Model:
The developed regression model, saved in a format that can be used for predictions.

Model Evaluation Report:
A report detailing the model's performance metrics (RMSE, R-squared, etc.) and evaluation results.

A/B Testing Results Report:
A report detailing the methodology, and results of the A/B testing, and the conclusions made from those results.

4. Communication & Presentation Phase:

Visualizations:
Charts, graphs, and other visual representations of the data and model results, tailored for the TLC executives.

Final Project Presentation:
A presentation summarizing the project's findings, model performance, and recommendations for the TLC.

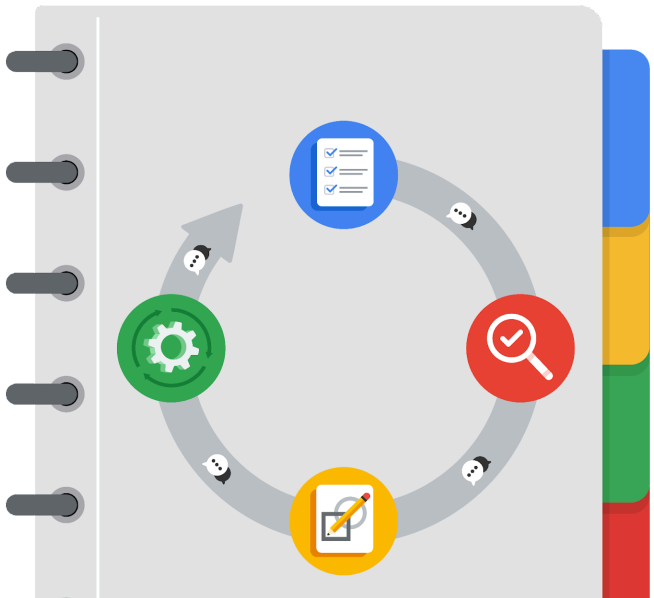
Documentation:
Documentation of the code, and methodology used in the project.

5. Potential Future Deliverables:

Fare Estimation App (Prototype or Final):
Depending on the project's scope, a working app that integrates the fare prediction model.

Deployment and Monitoring Plan:
A plan for deploying the model and monitoring its performance in a production environment.

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:** Execute

   Why did you select this stage for this task?

   This is done after the model is built, to see if it works well. It's the final step to check if we met our goals

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

   Why did you select these stages for this task?

   First, we analyze the data to come up with ideas (hypotheses). Then, we build tests to see if those ideas are right.

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

   Why did you select this stage for this task?

   This is the start of looking at the data to understand it, like checking what's inside the data.

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

   Why did you select these stages for this task?

   Exploring helps us find problems (analyze), and cleaning fixes those problems (construct).

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

   Why did you select this stage for this task?

   This is setting up how we'll do the whole project, which happens at the beginning.

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

   Why did you select this stage for this task?

   This is telling everyone what we found, which is the last step of the project.

7. **Compute descriptive statistics:** Analyze

   Why did you select this stage for this task?

   This is getting basic numbers about the data, like averages and counts, to understand it better.

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

   Why did you select these stages for this task?

   We make pictures of the data to see patterns (analyze) and to show our findings (construct).

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

   Why did you select this stage for this task?

   This is writing down what we're going to do, which is part of the planning.

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

Why did you select this stage for this task?

Building the model is creating the tool (construct), and checking how it works is analyzing it.

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

Why did you select this stage for this task?

This is gathering up the important stuff about the data to understand it.

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

Why did you select this stage for this task?

This is creating the machine learning tool.