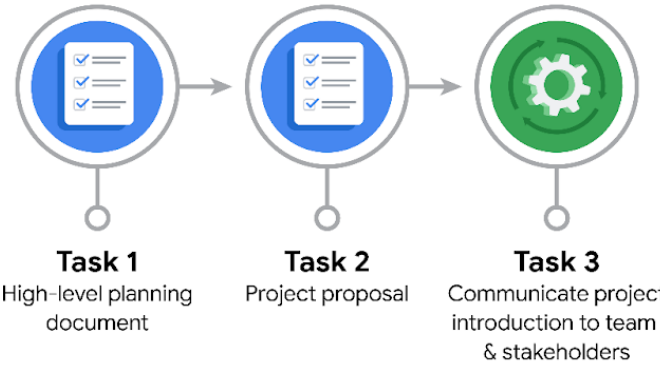
Foundations of Data Science

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 is the New York City Taxi and Limousine Commission.

- 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?

The primary problem this project aims to solve is the development of an accurate regression model for estimating taxi fares in advance for the New York City Taxi and Limousine Commission (TLC). By accomplishing this, several goals and anticipated impacts can be outlined: 1. **Improved Fare Estimates**: The core objective is to provide TLC with a reliable regression model that can estimate taxi fares accurately. This will directly benefit passengers by offering **transparency** in pricing and helping them plan their budgets more effectively.

2. **Data-Driven Decision-Making**: By leveraging historical data and advanced analytics, TLC will be able to make data-driven decisions regarding fare structures and pricing policies. This can lead to **optimized pricing strategies** that are responsive to market demands and competitive pressures.
3. **Operational Efficiency**: TLC can use the regression model to streamline operations, optimize resource allocation, and improve overall service efficiency. This could result in cost savings and a more effective allocation of resources.
4. **Revenue Growth:** An improved understanding of fare estimations and passenger preferences can potentially lead to revenue growth for TLC. Data-driven insights can help identify opportunities for upselling or cross-selling services.
5. **Stakeholder Satisfaction:** Satisfying the needs of both passengers and taxi service providers will foster a positive relationship between TLC and its stakeholders. This could lead to increased trust and support for TLC's initiatives.
6. **Competitive Advantage**: The ability to offer accurate fare estimations can give TLC a competitive edge in the transportation industry, attracting more passengers and service providers.

- What questions need to be asked or answered?

Project Initiation (Plan)

1. What are the specific objectives and goals of this project?
2. Who are the key stakeholders, and what are their expectations?
3. What is the scope of the project?
4. What resources (human, data, tools) are available for the project?

Data Acquisition (Acquire)

1. What data sources are available from the TLC, and what is the data's quality?
2. What data privacy and security measures need to be in place when accessing TLC's data?
3. Are there any legal or ethical considerations when accessing and using the data?

Data Preprocessing (Cleanse)

1. What data cleaning and preprocessing steps are necessary?
2. Which features or variables should be engineered or created for the regression model?

Exploratory Data Analysis (Explore)

1. What insights are we looking to gain from the exploratory data analysis (EDA)?
2. What visualizations and summary statistics will help convey key insights to stakeholders?

Model Development (Explore)

1. Which regression modeling technique is most suitable for this project?
2. How will the training and testing datasets be split, and what evaluation metrics will be used for model performance?

Model Validation (Explore)

1. What techniques will be used to validate the model's performance and reliability?
2. How will potential issues like overfitting or underfitting be addressed?

Insights and Reporting (Plan)

1. What insights and actionable recommendations will be derived from the model?
2. What format will the final report take, and how will the insights be presented to stakeholders?

Presentation to TLC (Plan)

1. When and how will the presentation to TLC be scheduled?
2. What key messages and visuals should be included in the presentation to effectively communicate findings?

Model Deployment (Plan)

1. How will the regression model be integrated into TLC's infrastructure?
2. What documentation and training materials will be provided to TLC for model usage and maintenance?

- What resources are required to complete this project?

Human Resources:

1. Data Analysts and Data Scientists: These professionals will play a central role in data preprocessing, exploratory data analysis, model development, and validation.
2. Project Manager: A project manager, such as the Senior Project Manager, Uli King, will oversee project planning, scheduling, resource allocation, and coordination among team members.
3. Data Analysis Manager: Deshawn Washington, in the role of Data Analysis Manager, will provide leadership, guidance, and supervision throughout the project.
4. Director of Data Analysis: Udo Bankole, as the Director of Data Analysis, will provide strategic oversight and ensure that the project aligns with the company's goals and standards.
5. Stakeholders from TLC: Engage with Juliana Soto (Finance and Administration Department Head) and Titus Nelson (Operations Manager) for input, feedback, and collaboration.
6. Communication and Presentation Specialists: Professionals skilled in communication and data presentation will be needed to prepare reports and presentations for TLC.

Technical Resources:

1. Data: Access to the TLC's historical taxi and limousine data, which includes trip records, fare information, and relevant attributes.
2. Computing Infrastructure: Sufficient computing power and cloud resources to support data analysis, model development, and deployment.
3. Data Analysis Tools: Utilize data analysis and machine learning tools such as Python libraries (e.g., Pandas, NumPy, Scikit-Learn) and visualization tools (e.g., Matplotlib, Seaborn).
4. Regression Modeling Software: Tools and libraries for regression model development (e.g., scikit-learn, TensorFlow, PyTorch).

5. Database Management: Database systems to store and manage large datasets efficiently.
6. Data Security and Compliance Tools: Ensure that data handling and storage meet security and compliance standards.

Additional Resources:

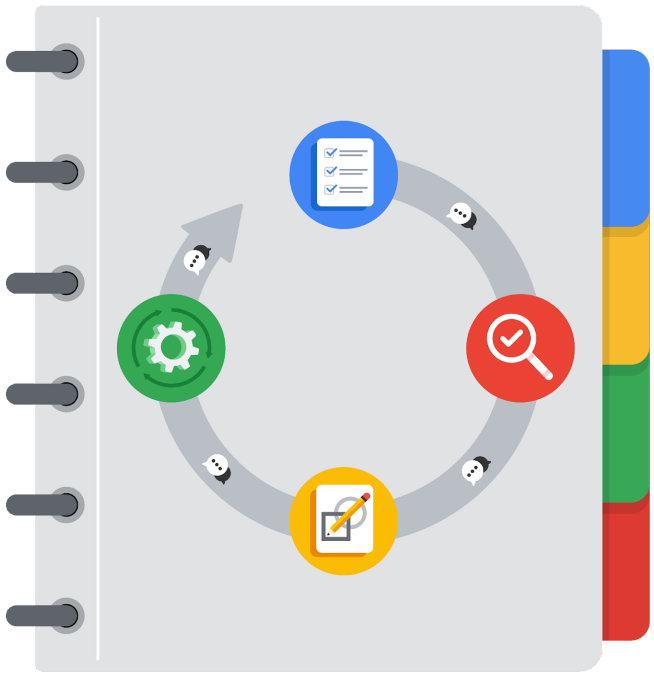
1. Documentation and Reporting Templates: Templates for documenting data processes, model development, and creating project reports.
2. Training Materials: Educational materials for user training, if required, to help TLC staff understand and use the regression model.
3. Meeting Facilities: Meeting rooms or virtual conferencing tools for project discussions, presentations, and coordination with stakeholders.
4. Project Management Software: Tools for project planning, tracking, and collaboration, such as Jira or Trello.
5. Communication Tools: Email, instant messaging, and video conferencing tools for efficient communication within the project team and with TLC stakeholders.
6. Data Privacy and Security Protocols: Guidelines and protocols for ensuring data privacy and security during data access and analysis.
7. External Expertise: Access to subject matter experts or consultants for specialized knowledge, if necessary.
8. Budget: Adequate funding for data acquisition, tool licensing, and any external resources required.

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

1. Project Proposal: The initial project proposal, outlining the project's scope, objectives, milestones, and stakeholders. This has already been created and submitted.
2. Data Acquisition Documentation: Documentation of the data sources accessed from TLC, including data dictionaries, metadata, and data quality assessments.
3. Data Preprocessing Report: A report detailing the steps taken to cleanse and preprocess the TLC dataset, addressing issues such as missing data, outliers, and feature engineering.
4. Exploratory Data Analysis (EDA) Findings: A summary of insights gained through EDA, including visualizations, summary statistics, and initial observations about the data.
5. Regression Model Development Report: Documentation of the regression model development process, including algorithm selection, data splitting, and model training.
6. Model Validation and Performance Report: A report on model validation results, including metrics for model performance evaluation and any adjustments made to improve the model's reliability.
7. Insights and Recommendations Report: A comprehensive report outlining key insights, trends, and actionable recommendations derived from the regression model.
8. Presentation to TLC: A presentation deck for TLC executives, summarizing the project's goals, findings, and recommendations, along with visual aids and data-driven insights.
9. Model Deployment Documentation: Documentation on how to deploy the regression model within TLC's infrastructure, including integration guidelines and any necessary code or configurations.
10. User Training Materials (if required): Educational materials, such as user manuals or training presentations, to facilitate TLC staff's understanding and use of the regression model.
11. Project Closure Report: A final report summarizing the project's overall outcomes, lessons learned, and any future recommendations or actions.

12. Meeting Notes: Detailed notes from project meetings, including discussions, decisions, and action items, to maintain a record of project progress.
13. Code Repository: A repository containing all code scripts and notebooks used in data analysis, model development, and deployment for transparency and future reference.
14. Data Security and Compliance Documentation: Documentation of data security and compliance measures implemented to protect sensitive information during the project.
15. Communication Records: Records of communication among project team members, including emails, instant messages, and meeting schedules.

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?

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

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

   Why did you select these stages for this task?

Analyze: Hypothesis testing is a statistical analysis technique used to evaluate data and draw conclusions about population parameters. In this task, you are actively analyzing the data to test hypotheses and make statistical inferences. It involves selecting appropriate statistical tests, calculating test statistics, and interpreting results. The primary focus here is on data analysis. Construct: The test is carried out.

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

   Why did you select this stage for this task?

In this stage, the focus is on understanding the data, its characteristics, and its potential insights. Starting the data exploration is the initial step in analyzing the dataset, which involves descriptive statistics, data visualization, and preliminary observations. It is a critical step in gaining insights and setting the foundation for further analysis and model construction.

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

   Why did you select these stages for this task?

Analyze (Data Exploration): The initial phase of data exploration, where you examine the dataset to understand its structure, characteristics, and patterns, falls under the "Analyze" stage. During this stage, you might perform basic data profiling, review summary statistics, and visualize the data to gain insights into its content and quality.

Plan: Planning takes place when you first make choices about the methods needed.

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

Why did you select this stage for this task?

This task involves creating the initial framework and structure for the project workflow. During the "Plan" stage, you define the overall project scope, objectives, milestones, and resources required. Establishing the structure for the project workflow is a foundational step in project planning, where you lay the groundwork for how the project will be organized, executed, and monitored. This task precedes the detailed analysis, construction, and execution phases of the project and sets the direction for subsequent actions.

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

Why did you select this stage for this task?

This task falls under the "Execute" stage of the PACE workflow because it involves the final presentation and communication of insights to both internal and external stakeholders. In this stage, the project team takes action by presenting the results, answering questions, and considering different viewpoints from stakeholders. It's the culmination of the project's analytical work and represents the final step in delivering value from the analysis to the organization or client.

7. **Compute descriptive statistics:** Analyze

Why did you select this stage for this task?

Computing descriptive statistics involves working with the data to summarize its key characteristics. This task falls under the Analyze stage of the PACE workflow because it is a fundamental step in understanding the data and its distribution. Descriptive statistics provide insights into the central tendency, variability, and distribution of data, helping analysts gain an initial understanding of the dataset. This analysis informs subsequent steps in the data analysis process, such as identifying outliers or trends during exploratory data analysis (EDA) and selecting appropriate modeling approaches during the Construct stage.

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

Why did you select these stages for this task?

Analyze Stage: Visualization begins with data assessment.

Construct Stage: While planning covers the initial groundwork for visualization creation, the actual construction of visualizations is where you design and build the charts, graphs, or dashboards based on the plan you developed. Therefore, this task also involves the Construct stage.

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

Why did you select this stage for this task?

During this stage, you define the scope of the project, identify goals, objectives, and stakeholder needs, and develop an overall project plan. The proposal serves as a roadmap for the entire project and sets the direction for subsequent tasks.

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

Why did you select this stage for this task?

Analyze: The analysis phase involves the actual construction of the regression model. This includes tasks such as data preprocessing, feature engineering, algorithm selection, model training, and hyperparameter tuning. The core of model building and analysis takes place in this stage, where you transform raw data into a functional predictive model.

Construct: The building of the regression model will take place in the construction phase.

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

Why did you select this stage for this task?

This task involves engaging with the data acquired from primary and secondary sources. It includes cleaning, reorganizing, and transforming the data, which are typical activities in the Analyze stage. Compiling summary information is an essential part of understanding the dataset's characteristics before proceeding to other stages.

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

Why did you select this stage for this task?

In this stage, you are actively constructing the predictive model by selecting an appropriate modeling approach, training the model, and refining it as needed. This task involves the construction of the core component of the project, which is the regression model for estimating taxi fares.