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.~~

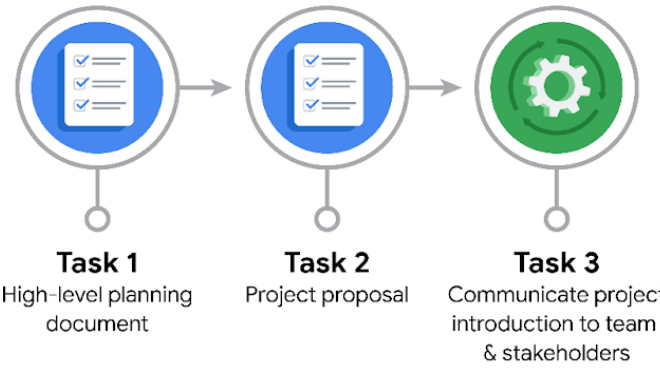
Relevant Interview Questions

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?

Data analytics team members, including Data Analysis Manager, Senior Data Analyst, Senior Project Manager, and Director of Data Analysis.

- 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 goal of the churn prediction project at Waze is to develop a machine learning model that accurately predicts user churn, specifically focusing on monthly user churn. This prediction model aims to:

- ☐ **Prevent User Churn**: By identifying users at risk of churning, Waze can proactively engage them with special offers, promotions, or improved services to retain them as satisfied and active users. This will reduce the number of users who uninstall the app or stop using it.
- ☐ **Improve User Retention**: Higher user retention rates indicate satisfied users who consistently use the Waze app over time. The prediction model will contribute to improving user retention by addressing the factors that lead to churn and implementing measures to retain users.
- ☐ **Boost Business Growth**: Retaining users and preventing churn is essential for business growth. Satisfied users are more likely to recommend the app to others, leading to increased user acquisition. Additionally, a lower churn rate means fewer resources spent on acquiring new users to replace those lost.
- ☐ **Identify Churn Factors**: The model will not only predict churn but also help identify the specific factors contributing to churn. This insight is valuable for Waze to make data-driven decisions and address issues such as user dissatisfaction or app usability.

The anticipated impact of this work on the larger needs of Waze's client, which includes users, advertisers, and partners, is significant. It will:

- ☐ **Enhance User Experience**: Preventing churn means that users continue to have access to Waze's navigation services, leading to improved customer satisfaction.
- ☐ **Optimize Advertising and Partnerships**: Satisfied and active users are more likely to engage with ads and partnerships within the app, creating more value for advertisers and partners.
- ☐ **Cost Reduction**: Preventing churn reduces the cost associated with acquiring new users, vehicle repair or replacement due to churn-related issues, and the potential impact on safety due to unexpected vehicle breakdowns.
- ☐ **Data-Driven Decision-Making**: The insights gained from the prediction model will enable Waze to make informed decisions about product development, marketing strategies, and user engagement initiatives.

Overall, this work aligns with the broader needs of Waze by contributing to its growth, user satisfaction, and the efficient use of resources, ultimately benefiting both the company and its user community.

- What questions need to be asked or answered?

Data-Related Questions:

1. **Data Availability**: *Is the necessary user data available and accessible for analysis and model development?*
2. **Data Quality**: *What steps will be taken to ensure the quality, cleanliness, and completeness of the data?*
3. **Data Privacy**: *How will we ensure compliance with data privacy regulations and maintain user data confidentiality?*
4. **Data Gathering**: *What is the plan for collecting and organizing relevant user data, including historical usage and churn data?*

Project Scope and Stakeholder Engagement:

5. **Project Objectives**: *What are the specific objectives and success criteria for this churn prediction project?*
6. **Stakeholder Identification**: *Who are the key stakeholders, both internal and external, and what are their expectations and requirements?*
7. **Communication Channels**: *How will we engage with stakeholders and leadership to align with their expectations and provide regular updates on project progress?*

Model Development and Evaluation:

8. **Churn Definition**: *How do we define churn in the context of this project? What constitutes a churned user?*
9. **Performance Metrics**: *What are the performance metrics we will use to evaluate the success of the churn prediction model?*
10. **Model Selection**: *What machine learning algorithms will be most suitable for predicting user churn?*
11. **Model Training**: *How will we split the data into training and testing sets for model development?*
12. **Model Evaluation**: *What criteria will be used to evaluate model performance, and how will we iteratively improve accuracy?*

Project Management:

13. **Project Timeline**: *What is the expected timeline for the project, including milestones and key deadlines?*
14. **Team Roles**: *What are the roles and responsibilities of each team member within the data analytics team?*
15. **Infrastructure Setup**: *How will we set up data storage and management infrastructure to support the project?*
16. **Data Update Frequency**: *How often will real-time data updates be integrated into the churn prediction model?*
17. **Regular Model Updates**: *What schedule will be established for model updates to ensure it reflects changing data patterns?*

These questions serve as a foundation for effective project planning and execution, ensuring that the team has a clear understanding of project goals, data requirements, and stakeholder expectations. Answering these questions will help guide the team in achieving the project's objectives and delivering valuable insights to Waze leadership.

- What resources are required to complete this project?

To successfully complete the churn prediction project at Waze, several key resources are required. These resources encompass data, tools, personnel, and infrastructure. Here is a breakdown of the essential resources:

1. **Data Sources**: Access to a comprehensive dataset of historical user interactions and churn data is fundamental. This includes data on user activities, app usage, feedback, and reasons for churn.
2. **Data Quality Assurance**: Resources and processes for data cleaning and quality assurance are crucial to ensure that the dataset is accurate, complete, and free from errors.
3. **Machine Learning Tools**: Tools and software for machine learning model development, including libraries like scikit-learn and TensorFlow, are essential.
4. **Data Storage and Management**: Infrastructure for storing and managing large datasets efficiently, such as cloud-based data storage solutions (e.g., AWS S3, Google Cloud Storage) or on-premises servers.
5. **Data Analysis Tools**: Tools for data analysis and exploration, including data visualization software (e.g., Tableau, Matplotlib) and data processing tools (e.g., Pandas).
6. **Model Training Resources**: Computational resources, including CPUs or GPUs, for training machine learning models.
7. **Personnel**: Skilled data professionals, including data analysts, data scientists, and project managers, are required to execute the project tasks effectively.
8. **Project Management Software**: Tools for project planning, tracking, and collaboration, such as project management platforms (e.g., Jira, Asana) and communication tools (e.g., Slack, Microsoft Teams).
9. **Data Privacy and Compliance**: Resources for ensuring data privacy and compliance with relevant regulations, including legal counsel if necessary.
10. **Communication and Reporting Tools**: Tools for communication within the team and reporting progress to stakeholders, including email, presentation software (e.g., PowerPoint), and data visualization tools.
11. **Infrastructure for Real-Time Data**: Infrastructure for collecting and integrating real-time data from the Waze app, including servers, APIs, and data pipelines.
12. **Model Deployment Environment**: Infrastructure and systems for deploying the churn prediction model in a production environment, including cloud-based deployment platforms (e.g., AWS SageMaker) or on-premises servers.
13. **Regular Model Update Schedule**: Resources and processes for maintaining and updating the model as new data becomes available.
14. **Training and Skill Development**: Resources for ongoing training and skill development for team members to stay updated with the latest data analytics and machine learning techniques.
15. **Feedback Mechanism**: A feedback mechanism to gather insights and feedback from users and stakeholders, helping improve the model over time.
16. **Documentation and Knowledge Sharing**: Tools and processes for documenting project activities, findings, and model deployment procedures to facilitate knowledge sharing within the team and with future stakeholders.

The availability and effective utilization of these resources will play a crucial role in the success of the churn prediction project, enabling the team to achieve its goals of improving user retention and fostering business growth.

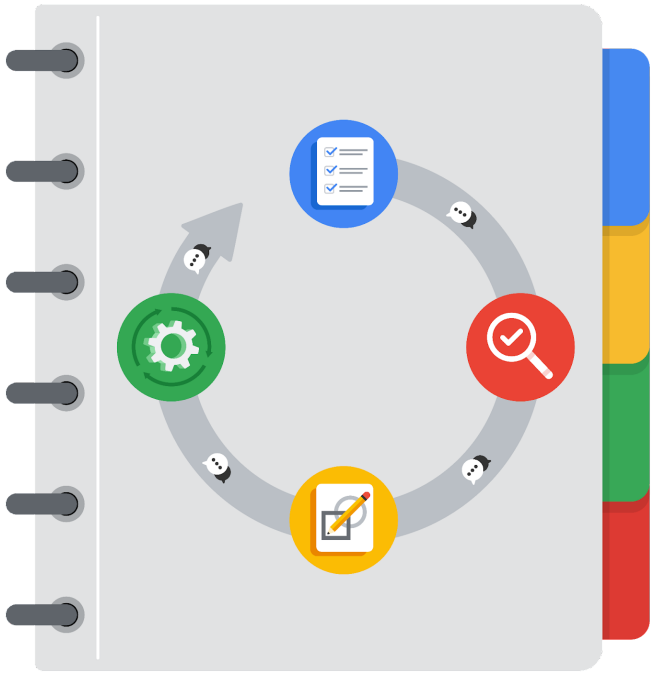
- What are the deliverables that will need to be created over the course of this project?

Throughout the course of the churn prediction project at Waze, several deliverables will need to be created to ensure the project's success and to communicate findings and progress effectively. Here are the key deliverables:

1. **Project Proposal**: This initial document outlines the project's objectives, milestones, and overall strategy. It defines the scope, key questions, and considerations, providing a high-level overview of the project.
2. **Data Collection and Preparation**: Deliverables in this category include:
- **Data Gathering Plan**: Document detailing sources and methods for collecting user data.
- Data Cleaning and Preprocessing Report: Documentation of data cleaning steps, transformations, and handling of missing values.
- **Exploratory Data Analysis (EDA) Report**: Insights gained from EDA, visualizations, and data summaries.
3. **Model Development**: Key deliverables related to constructing the churn prediction model include:
- **Machine Learning Model**: The developed predictive model, ready for deployment.
- **Model Evaluation Report**: Analysis of model performance, including metrics and visualizations.
- **Model Fine-Tuning Documentation**: Details on any adjustments made to improve model accuracy.
4. **Real-Time Data Integration**: As part of the execution phase, these deliverables are essential:
- **Real-Time Data Integration Pipeline**: Documentation and code for collecting and integrating real-time user data.
- **Deployed Model**: The churn prediction model integrated into Waze's production environment.
- **Schedule for Model Updates**: A plan for regularly updating the model to adapt to changing data patterns.
5. **Communication and Reporting**: These deliverables keep stakeholders informed:
- **Progress Reports**: Regular updates on project milestones, challenges, and achievements.
- **Final Presentation**: A comprehensive presentation summarizing the project's results, including model performance, insights, and recommendations.
- **Documentation**: Comprehensive documentation of all project stages, including code, data, and methodology.
6. **Documentation and Code Repository**: A centralized repository for storing project-related documents, code, and datasets, ensuring transparency and future reference.
7. **Stakeholder Engagement Plan**: A plan outlining how the team will engage with stakeholders, leadership, and other relevant parties throughout the project.
8. **Data Privacy and Compliance Documentation**: Records of measures taken to ensure data privacy and compliance with relevant regulations.

These deliverables collectively support effective project management, data analysis, model development, and communication with stakeholders. They ensure that the project stays on track, meets its objectives, and provides valuable insights to prevent user churn on the Waze app.

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?

While model evaluation typically occurs during the "Analyzing" phase, it's also relevant to discuss it in the "Executing" phase within the context of deploying and implementing the model in a production environment.

Implement real-time data collection from app users' interactions: In this phase, we are actively collecting real-time data from users' interactions with the Waze app. This data includes information about user behavior and usage patterns.

Develop data pipelines for continuous data updates: Data pipelines are essential for processing and updating the data continuously. As we receive new data from user interactions, it needs to be integrated seamlessly into our existing dataset.

Integrate real-time data into the churn prediction model: During the execution phase, we are not only collecting real-time data but also integrating it into the churn prediction model. This involves ensuring that the model is up to date with the latest user behavior.

Deploy the churn prediction model in a production environment: When we deploy the model in a real-world, production environment, it starts making predictions based on the real-time data. The model's performance in this live environment needs to be monitored.

Establish a schedule for regular model updates to reflect changing data patterns: In the execution phase, we set up processes for periodic updates to the model to adapt to changing data patterns. These updates may include retraining the model or adjusting its parameters.

The reason for discussing model evaluation in the "*Executing*" phase is to emphasize that, during model deployment and operation in a live environment, we continue to assess its performance and make necessary adjustments to ensure its effectiveness. This ongoing evaluation and adaptation are crucial to the success of the churn prediction system in practice.

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

Why did you select these stages for this task?

I selected "*Analyze*" and "*Construct*" as the stages for conducting hypothesis testing because these stages align with the typical workflow of hypothesis testing in data analysis. Here's why each stage is relevant:

1. **Analyze (A)**: This stage involves data exploration and understanding. Before conducting hypothesis testing, it's crucial to analyze the dataset to identify patterns, relationships, and potential factors that might be relevant to the hypothesis. During this phase, data analysts often perform descriptive statistics, data visualization, and exploratory data analysis (EDA) to get insights into the data's distribution and characteristics. This initial analysis helps in formulating hypotheses and understanding the data's context.

2. **Construct (C)**: In the "*Construct*" stage, data analysts design and execute statistical tests to test specific hypotheses. This phase involves selecting the appropriate statistical test based on the type of data and the research question. It includes setting up null and alternative hypotheses, choosing a significance level (alpha), conducting the test, and interpreting the results. This is the phase where the actual hypothesis testing takes place, and conclusions are drawn based on statistical evidence.

By including "*Analyze*" and "*Construct*" within the PACE framework for hypothesis testing, it ensures that data analysts follow a structured approach from data exploration and hypothesis formulation (Analyze) to the actual testing and interpretation (Construct). This helps in maintaining a clear and organized workflow during hypothesis testing, which is essential for rigorous and reliable data analysis.

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

Why did you select this stage for this task?

The choice to begin exploring the data during the "*Analyze*" phase is a common practice in data science and analytics projects. Here are the reasons behind this choice:

1. **Understanding Data**: Exploratory Data Analysis (EDA) allows data analysts to gain a deep understanding of the dataset. This includes understanding the structure, distributions, and characteristics of the data. It helps identify potential challenges and patterns within the dataset.

2. **Data Cleaning**: EDA often reveals data quality issues such as missing values, outliers, or inconsistencies. Addressing these issues early in the project is crucial to ensure the data's integrity and reliability.

3. **Feature Selection**: During EDA, analysts can identify which features (variables) in the dataset are likely to be relevant for the prediction task. This is important for selecting the right set of features for model development, which can have a significant impact on model performance.

4. **Initial Insights**: EDA can provide initial insights into user behavior and patterns that may be valuable for later stages of the project. These insights can guide the selection of appropriate machine learning algorithms and the formulation of hypotheses.

5. **Data Visualization**: EDA often involves creating data visualizations such as histograms, scatter plots, and heatmaps. These visualizations help in conveying complex information in a more understandable and interpretable manner, facilitating discussions with stakeholders.

6. **Hypothesis Generation**: Through EDA, analysts may form hypotheses about factors that could influence user churn. These hypotheses can guide further analysis and experimentation in later stages of the project.

In summary, the "*Analyze*" phase, specifically EDA, serves as a crucial foundation for the entire data analytics project. It helps ensure data quality, provides initial insights, and informs decisions about feature selection and model development, making it a logical starting point for exploring the data.

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

Why did you select these stages for this task?

The "*Plan*" stage sets the groundwork for effective data exploration and cleaning by defining objectives, stakeholders, data collection processes, and infrastructure, while the "*Analyze*" stage focuses on the actual data cleaning and exploration tasks, including preprocessing, EDA, feature identification, and metric definition. This structured approach ensures that the data is prepared effectively for subsequent phases of the project.

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

Why did you select this stage for this task?

I selected "*Plan*" because it provides a structured and organized approach to project management and execution. The PACE framework is particularly relevant for projects in the field of data analytics and data science, where clear planning, analysis, construction, and execution are essential for success. Here's why I chose it:

1. Structured Approach: The PACE framework divides the project into distinct phases, making it easier to manage and ensuring that each phase is well-defined and executed.

2. Clear Milestones: It encourages the establishment of clear milestones and objectives for each phase, helping the team stay on track and ensuring that progress can be measured effectively.

3. Flexibility: While it provides structure, the PACE framework is flexible enough to accommodate changes and adjustments as the project progresses, which is crucial in dynamic fields like data analytics.

4. Alignment with Project Goals: PACE aligns well with the project's goals of developing a machine learning model for user churn prediction, as it emphasizes planning, analysis, construction, and execution – all of which are critical for model development.

5. Efficient Communication: The framework promotes effective communication among team members and stakeholders by providing a common structure and language for discussing project progress.

Overall, the PACE framework is a valuable tool for organizing and managing complex projects, such as the churn prediction project at Waze, and it ensures that each phase of the project is approached systematically and strategically.

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

Why did you select this stage for this task?

I selected "*Executing*" as the stage for "*Communicate final insights with stakeholders*" because the "*Executing*" phase in the PACE framework encompasses the deployment and implementation of the project's results into a practical environment. Communicating final insights with stakeholders is a crucial part of this phase because it involves taking the actionable insights derived from the analysis and model development and putting them into practice to achieve the project's goals. Here's why this choice aligns with the PACE framework:

1. **Deployment of Insights**: In the "Executing" phase, the focus is on deploying the churn prediction model and integrating real-time data. As part of this deployment, the final insights derived from the model need to be communicated to stakeholders. This may involve implementing strategies or actions based on those insights.

2. **Practical Application**: Communicating insights is not just about sharing information; it's about putting that information to practical use. This could involve alerting specific teams or individuals when high-risk users are identified or initiating targeted retention strategies based on the insights.

3. **Iterative Process**: The "*Executing*" phase often involves ongoing monitoring and adjustments. As insights are communicated and actions are taken, the project team and stakeholders need to continue assessing the impact and making improvements, which is in line with the iterative nature of this phase.

While communication with stakeholders can happen throughout the project, the final communication of insights and the practical implementation of those insights typically occur during the "Executing" phase when the project is transitioning from development to deployment and real-world use.

7. **Compute descriptive statistics:** Analyze

Why did you select this stage for this task?

I selected "*Analyze*" for "*Compute descriptive statistics*" because computing descriptive statistics is an analytical task that involves examining data to summarize its main characteristics. Descriptive statistics provide insights into the central tendency, dispersion, and distribution of data, which are essential for understanding and interpreting the data effectively. This task typically involves calculating measures such as mean, median, standard deviation, and percentiles, among others.

In the context of the churn prediction project at Waze, the analysis of user data is a critical step. Understanding the distribution of user behavior, such as how frequently users engage with the app, the average time spent, and other relevant metrics, is essential for building an accurate churn prediction model. Descriptive statistics help uncover patterns and trends in the data, which can guide feature selection and model development.

In summary, choosing "*Analyze*" for "*Compute descriptive statistics*" aligns with the analytical nature of the task and its importance in the data analysis process, particularly in the context of a project like churn prediction where data understanding is fundamental.

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

Why did you select these stages for this task?

I selected "Analyze" and "Construct" stages for visualization building because they are typically the stages in which data analysts and data scientists plan and create visualizations to communicate insights from the data effectively. Here's why each stage is important:

Analyze (A):

Exploratory Data Analysis (EDA): This stage involves exploring the dataset to understand its structure, identify patterns, outliers, and relationships between variables. Visualizations, such as histograms, scatter plots, and box plots, play a crucial role in EDA to provide initial insights into the data.

Visualization Planning: During the analysis phase, analysts plan what visualizations will best convey the key findings and insights uncovered during EDA. They decide which charts, graphs, or dashboards will be most effective in presenting the data.

Construct (C):

Building Visualizations: In the construction stage, analysts and data scientists create the actual visualizations based on their EDA findings and the visualization plan. This involves using tools like Python libraries (matplotlib, seaborn), R, or specialized data visualization tools (Tableau, Power BI) to generate charts and graphs.

Iterative Improvement: The construction phase allows for iterative improvement of visualizations. Analysts can experiment with different chart types, colors, and styles to optimize the clarity and impact of the visuals.

Integration with Data Analysis: Visualizations constructed in this stage are closely tied to the data analysis process. They are used to illustrate trends, patterns, and insights uncovered during data analysis, making them an integral part of the storytelling process.

By incorporating visualization tasks into both the "Analyze" and "Construct" stages, it ensures that visualizations are not an afterthought but are strategically planned and built to support data analysis and communicate findings effectively to stakeholders.

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

Why did you select this stage for this task?

I selected "Plan" for writing a project proposal because the act of planning is essential before embarking on any significant project. In this scenario, May Santner, the Data Analysis Manager at Waze, asked for a project proposal to outline the plan of action for the team's churn prediction project. This proposal serves as the initial step in organizing and communicating the project's goals, objectives, milestones, and strategies.

The planning phase is crucial because it helps establish a clear roadmap for the project, ensuring that all team members are aligned with the project's scope, objectives, and expectations. It also helps identify potential challenges and risks early on, allowing the team to address them proactively. Additionally, planning sets the foundation for effective project execution, monitoring, and success measurement.

In summary, selecting "Plan" for writing a project proposal aligns with the PACE (Planning, Analyzing, Constructing, Executing) framework, which emphasizes the importance of thorough planning before diving into data analysis and project execution.

10. Build a regression model: Analyze and Construct

Why did you select this stage for this task?

The choice to "Analyze" and "Construct" in the PACE framework for building a regression model is based on the specific stages involved in developing and implementing such a model. Let me explain why these stages were selected:

1. **Analyze (A)**:

- **Data Exploration**: Before constructing a regression model, it's crucial to thoroughly understand the data. This involves exploring the dataset, identifying potential outliers, understanding the distribution of variables, and gaining insights into the relationships between predictor variables and the target variable (which is typically the one being regressed upon). EDA helps in feature selection and understanding the data's characteristics, which are critical for model building.

- **Data Preprocessing**: Data cleaning and preprocessing are fundamental to regression modeling. This includes handling missing values, scaling or normalizing features, and addressing multicollinearity (correlations between predictors). Preprocessing ensures that the data is in a suitable form for modeling and helps in reducing noise in the model.

- **Feature Selection**: Analyzing the importance of different features in explaining the target variable is essential. Feature selection techniques, such as examining feature correlations, can be applied in the analysis phase to determine which features should be included in the regression model.

2. **Construct (C)**:

- **Model Selection**: Once the data has been analyzed, and relevant features have been identified, the construction phase involves selecting an appropriate regression model. This includes choosing between different types of regression models, such as linear regression, multiple regression, or more complex models like ridge or lasso regression. The choice of model depends on the nature of the data and the assumptions of the regression method.

- **Model Building**: In this phase, the selected regression model is built using the cleaned and preprocessed data. This includes estimating the model coefficients and parameters, such as the intercept and slope coefficients in the case of linear regression. The construction phase is where the model "learns" from the data.

- **Model Evaluation**: After constructing the model, it's essential to evaluate its performance to ensure its effectiveness in predicting the target variable. This involves using various evaluation metrics, such as Mean Squared Error (MSE) or R-squared, to assess how well the model fits the data. If the model doesn't perform well, adjustments may be needed in the construction phase.

By incorporating the "Analyze" and "Construct" stages, the PACE framework ensures a structured approach to regression modeling that includes data exploration, preprocessing, and model building. These phases are interconnected and vital for developing an accurate and effective regression model.

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

Why did you select this stage for this task?

In the provided scenario for the Waze churn prediction project, the task to "*Compile summary information about the data*" falls under the "*Analyze*" stage of the PACE framework. Here's why:

1. **Nature of the Task**: The task involves the examination and compilation of summary information regarding the user data that will be used for the churn prediction project. This action primarily pertains to gaining insights from the data, understanding its characteristics, and summarizing key statistics and attributes.

2. **Data Exploration**: Analyzing and summarizing data is a fundamental step in the data analytics process. It helps the team become familiar with the dataset, identify potential challenges, and make informed decisions about data preprocessing and feature selection.

3. **EDA (Exploratory Data Analysis)**: Compiling summary information is often a crucial part of EDA. EDA aims to discover patterns, trends, and anomalies within the data. Summarizing data through descriptive statistics, visualizations, and statistical measures is essential for EDA.

4. **Data Understanding**: This task contributes to a deeper understanding of the data, which is essential for subsequent stages like feature engineering and model selection. It ensures that the team is well-informed about the dataset's structure and content.

5. **Foundation for Decision-Making**: The summary information collected during this task serves as the foundation for making informed decisions about feature selection, model choice, and the overall analytical approach. It guides the team in defining relevant metrics and KPIs.

In summary, the task "*Compile summary information about the data*" involves activities related to data exploration and understanding, which are integral parts of the "*Analyze*" stage in the PACE framework for data analytics projects.

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

Why did you select this stage for this task?

I selected "*Construct*" for building the machine learning model because it aligns with the PACE framework, which stands for Planning, Analyzing, Constructing, and Executing. In this context:

1. **Planning (P)**: In the planning phase, we define project objectives, scope, and success criteria. We also identify key stakeholders and establish communication channels. It sets the foundation for the project.

2. **Analyzing (A)**: In the analyzing phase, we clean and preprocess data, perform exploratory data analysis (EDA), and identify relevant features. This phase focuses on understanding the data and preparing it for modeling.

3. **Constructing (C)**: The constructing phase is where the actual building or construction of the machine learning model takes place. This is where we select appropriate machine learning algorithms, split the data for training, train and fine-tune the model, and evaluate its performance. It's the core phase of model development.

4. **Executing (E)**: Finally, in the executing phase, we implement the model in a production environment and establish processes for continuous updates and real-time data integration.

So, "*Construct*" specifically refers to the stage where we build and develop the machine learning model, making it a fitting choice in the context of the PACE framework and the project's workflow.