Course Two
Get Started with Python

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 questions in the Course 2 PACE strategy document
- ☐ Answer the questions in the Jupyter notebook project file
- ☐ Complete coding prep work on project's Jupyter notebook
- ☐ Summarize the column Dtypes
- ☐ Communicate important findings in the form of an executive summary

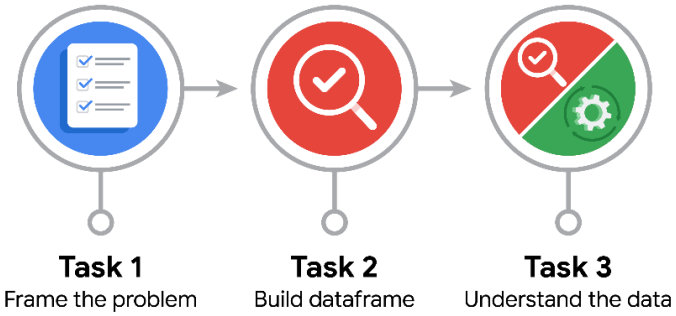
Relevant Interview Questions

Completing the end-of-course project will help you respond these types of questions that are often asked during the interview process:

- Describe the steps you would take to clean and transform an unstructured data set.
- What specific things might you look for as part of your cleaning process?
- What are some of the outliers, anomalies, or unusual things you might look for in the data cleaning process that might impact analyses or ability to create insights?

Reference Guide

This project has three tasks; the visual below identifies how the stages of PACE are incorporated across those tasks.



Data Project Questions & Considerations

PACE: Plan Stage

- How can you best prepare to understand and organize the provided information?

1. **Data Documentation**: Reviewing any available documentation or data dictionaries that describe the structure of the data, the meaning of each column, and any potential data quality issues. Understanding the data source and its characteristics is crucial.

2. **Data Cleaning**: Performing data cleaning tasks to address missing values, duplicate records, and outliers to ensure that the data is accurate and reliable for analysis.

3. **Data Exploration**: Use tools like pandas for Python or other data analysis software to explore the data, including examining summary statistics, data distributions, and visualizations to identify patterns and trends.

4. **Feature Engineering**: The need to create new features or transform existing ones to better capture the information relevant to predicting churn may occur based on observations from previous steps.

5. **Data Sampling**: Considering taking a random sample to speed up the initial exploratory data analysis. This can be particularly useful for understanding the data's structure before analyzing the entire dataset.

6. **Data Visualization**: Creating visualizations such as histograms, scatter plots, and correlation matrices to gain insights into the relationships between variables and potential factors contributing to churn.

7. **Data Preparation**: Format the data in a way that is suitable for modeling. This may include encoding categorical variables, scaling numerical features, and splitting the data into training and testing sets.

8. **Documentation**: Documenting all the data preparation steps and observations. This documentation will be valuable for sharing insights with team members and stakeholders.

9. **Collaboration**: Collaborating with other team members, including data scientists, domain experts, and business analysts, to gather their insights and perspectives on the data. Their input can help refine the analysis.

10. **Quality Control**: Double-checking the data for accuracy and consistency, and verifying that it aligns with the project's objectives and requirements.

- What follow-along and self-review codebooks will help you perform this work?

yes

- What are some additional activities a resourceful learner would perform before starting to code?

Read the docs.

PACE: Analyze Stage

- Will the available information be sufficient to achieve the goal based on your intuition and the analysis of the variables?

Yes data looks clean and insightful towards churn prediction.

- How would you build summary dataframe statistics and assess the min and max range of the data?

Using the describe function of a pandas dataframe to produce statistics and using min and max function with axis set to 1 to get the min and max values for each column of the dataframe.

- Do the averages of any of the data variables look unusual? Can you describe the interval data?

The median user in both the churned and retained groups drove approximately 73 kilometers per drive. However, when we examine their driving patterns further, we observe significant differences. Churned users drove an average of 608.77 kilometers per day they were actively driving, which is considerably higher than the 247.47 kilometers per day for retained users. Additionally, churned users had a much higher frequency of driving, averaging 8.33 trips per driving day, while retained users averaged 3.35 trips per driving day. These findings highlight substantial distinctions in driving behavior between the two user groups.

PACE: Construct Stage

Note: The Construct stage does not apply to this workflow. The PACE framework can be adapted to fit the specific requirements of any project.

PACE: Execute Stage

- Given your current knowledge of the data, what would you initially recommend to your manager to investigate further prior to performing exploratory data analysis?

It's crucial to understand the methodology behind data collection and the sampling process. Investigate how the user data was collected, whether it represents a random sample, and if any biases or limitations may exist.

Perform data quality checks to identify missing values, duplicates, outliers, or inconsistencies in the dataset. Ensure that data cleaning procedures are in place to handle any data issues.

Clarify the definitions and meanings of variables in the dataset, especially those related to user behavior and churn. Make sure everyone on the team has a consistent understanding of these definitions.

Ensure that the collection and use of user data comply with privacy regulations and ethical standards. Confirm that appropriate consent and data protection measures are in place.

Investigate whether there is any sampling bias in the dataset, particularly regarding user demographics or geographic regions. Assess whether the dataset represents a diverse user base.

Understand the context in which the data was collected. This includes understanding the time period, market conditions, and any external events that may have influenced user behavior.

Create comprehensive documentation that includes metadata, variable descriptions, and any data transformations or preprocessing steps applied to the dataset.

Gather input from stakeholders, including domain experts and business analysts, to identify specific questions or hypotheses that should be addressed during exploratory data analysis.

- What data initially presents as containing anomalies?

`driven_km_drives`, `driving_days` and `drives`.

- What additional types of data could strengthen this dataset?

Demographic Data: Including demographic information such as age, gender, income, and location can provide insights into how different user segments behave and churn. This data can help identify

whether certain demographics are more prone to churn.

User Feedback and Surveys: Gathering feedback from users through surveys or reviews can provide qualitative data about their experiences, pain points, and reasons for churn. This can complement the quantitative data in the dataset.

App Usage Metrics: Collecting more granular data on how users interact with the app, such as specific features used, frequency of app usage, and session duration, can offer a deeper understanding of user engagement.

Customer Support Interactions: Data on customer support interactions, including inquiries, complaints, and resolutions, can shed light on user satisfaction and the effectiveness of support in reducing churn.

Competitor Data: Information about users' interactions with competing navigation apps or services can help identify competitive factors influencing churn.

User Behavior Trends Over Time: Collecting data over an extended period allows for the analysis of user behavior trends, seasonality, and changes in behavior patterns leading up to churn events.

Location and Route Data: Detailed location and route information can provide insights into the types of journeys users undertake and their preferences for different routes or destinations.

App Version and Updates: Tracking users' app versions and responses to updates can help determine whether app changes or updates have an impact on churn rates.

User Sentiment Analysis: Analyzing user-generated content such as reviews, social media comments, or app store ratings using sentiment analysis can reveal user sentiment and satisfaction levels.

Customer Lifecycle Data: Understanding where users are in their customer lifecycle (e.g., new users, active users, lapsed users) can help tailor retention strategies for different user segments.

In-App Events and Triggers: Monitoring in-app events, such as the creation of user profiles, setting preferences, or completing specific actions, can provide insights into user engagement and potential churn triggers.

External Data Sources: Incorporating external data sources, such as weather data, traffic conditions, or local events, can help contextualize user behavior and identify external factors affecting churn.