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

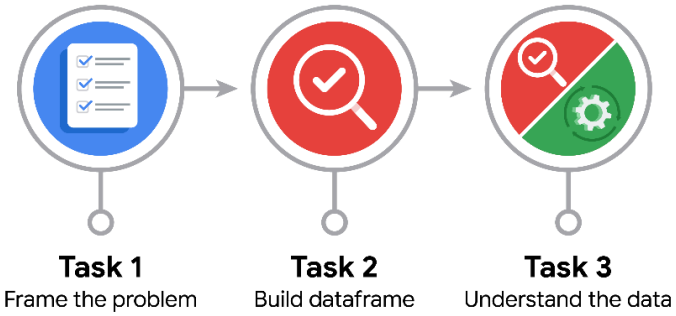
Relevant Interview Questions

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?

To prepare for the project, I thoroughly read emails from Rosie Mae Bradshaw and Orion Rainier, identified vital objectives, and understood the nature of claims and opinions in the provided datasets. Also, I Reviewed team structures, referring to the project proposal completed in Course 1 for additional context. I familiarized myself with any provided guidelines or templates, considering the audience and purpose of the tasks. Additionally, I broke down coding and data organization tasks into actionable steps, ensured I had the necessary tools, and Took notes on key points and questions to stay organized during the project. This comprehensive approach will enhance my understanding and organization of the provided information.

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

The "Course 2 TikTok Project Lab" is an invaluable follow-along codebook to facilitate coding and data preparation. This resource guides me through essential Python tasks, offering practical insights for effectively importing, inspecting, and organizing the data. The "Course 2 PACE Strategy Document" and "Executive Summary Templates" act as structured frameworks for self-review and strategic

planning. Though they do not contain actual code, these codebooks provide valuable guidelines for documenting my project strategy, findings, and recommendations. Utilizing these resources ensures a comprehensive understanding of the tasks at hand and effective organization and communication of my work, aligning seamlessly with the project objectives.

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

Before delving into coding, a resourceful learner undertakes several key activities to ensure a comprehensive understanding of the project. This includes meticulously reviewing the provided dataset(s) to grasp its structure and potential challenges. Researching relevant Python libraries, such as pandas and numpy, is essential for effective data manipulation. Exploring similar projects or case studies helps to glean insights into common approaches and best practices. Seeking clarification on any ambiguities in the project requirements and reviewing team communications provide valuable context. Ensuring the coding environment is set up, creating a task checklist, and drafting a coding plan contribute to an organized approach. Additionally, considering potential data visualization techniques and gathering coding references further solidifies the learner's readiness for the coding phase.

PACE: Analyze Stage

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

The PACE (Prepare, Analyze, Communicate, Execute) framework is a strategic approach to project management. In the Analyze stage, assessing whether the available information is sufficient to achieve the project goal is crucial. Intuition plays a role, but thoroughly analyzing the variables involved should supplement it. Evaluate the dataset's comprehensiveness, the relevance of the variables, and potential gaps in information. If uncertainties or data limitations are identified during this stage, it may be necessary to revisit the preparation phase to gather additional data or refine the project scope. Regularly reassessing the sufficiency of information ensures a robust foundation for the subsequent stages of the project.

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

To build summary data frame statistics and assess the minimum and maximum range of the data, I would utilize Panda's library in Python. The `describe()` function generates comprehensive summary statistics for each numeric column in the DataFrame, offering insights into the data's central tendency, dispersion, and distribution. This includes values such as mean, standard deviation, and quartiles. Simultaneously, the `min()` and `max()` functions extract each column's minimum and maximum values, providing a clear understanding of the data range. By examining these summary statistics, I can quickly identify critical characteristics of the dataset, enabling informed decision-making and further exploration in subsequent stages of my data analysis or machine learning project.

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

To evaluate whether the averages of data variables appear unusual, it is essential to compare them against expected norms or industry benchmarks. Unusual averages might indicate outliers, errors, or unique patterns within the dataset. Visualizations, such as histograms or box plots, can help identify the distribution of values, providing insights into potential anomalies or trends. For interval data, which consists of numeric values with equal intervals, describing the data involves examining central tendency measures like the mean and median, along with dispersion measures such as the range and standard deviation. Understanding the variability within the intervals allows for a comprehensive characterization of the data's distribution. This assessment is inherently context-specific, and a thorough exploration of the dataset's characteristics is crucial for interpreting the significance of any observed patterns or deviations in averages and interval data.

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?

Before delving into exploratory data analysis (EDA) in the execution stage of the PACE framework, I recommend a comprehensive investigation into several vital aspects. Firstly, a meticulous examination of data quality is imperative to identify and rectify any missing values, outliers, or inconsistencies that could impact the reliability of subsequent analyses. Prioritizing and understanding the significance of variables based on their relevance to the project goal helps focus the EDA on crucial aspects. Leveraging domain-specific knowledge and consulting experts provides deeper insights while analyzing data distributions, exploring correlations, and assessing the need for data transformations to ensure a nuanced dataset understanding. Formulating initial hypotheses based on observed patterns or domain expertise guides the EDA towards specific questions, fostering a more purposeful exploration. Additionally, ensuring compliance with data privacy regulations and ethical standards is crucial to maintaining trust and integrity throughout the analysis process. These preliminary investigations set the stage for a well-informed and targeted exploratory data analysis, facilitating the extraction of meaningful insights.

- What data initially presents as containing anomalies?

Identifying anomalies in data often revolves around specific characteristics that deviate from the norm, and several types of data may initially signal potential anomalies. Numeric variables exhibiting

extreme values or outliers may hint at anomalies impacting statistical measures and overall analysis. High incidences of missing values in certain variables suggest anomalies or issues in data collection, urging a closer examination. Inconsistencies in categorical variables, such as unexpected or irregular categories, could point to errors in data entry or coding, potentially affecting subsequent analyses. Time-related data may reveal anomalies like sudden spikes or drops, indicating specific events or issues in data collection over time. Unexpected relationships or correlations between variables could signify anomalies, prompting further investigation into underlying causes.

Additionally, data entry errors, including typos or inaccuracies, and data drift—gradual changes in distribution over time—may represent potential anomalies that require scrutiny. Leveraging visualization, statistical methods, and domain knowledge is paramount to accurately pinpointing and addressing these anomalies. Regular updates to anomaly detection methods and collaboration with subject matter experts enhance the effectiveness of anomaly identification processes.

- What additional types of data could strengthen this dataset?

Several types of additional data could be incorporated to fortify and enrich the existing dataset, providing a more comprehensive and nuanced understanding of the phenomena under investigation. Demographic data, encompassing factors such as age, gender, and location, would offer valuable insights into the population's characteristics. Including temporal data, including timestamps and date information, would introduce a temporal dimension, enabling analyses of trends, patterns, and changes over time. External economic indicators could be integrated to contextualize the observed data within the broader economic landscape. User engagement metrics like click-through rates and time spent on a platform shed light on user behavior and interaction patterns. Geospatial data would allow the exploration of spatial relationships, geographic patterns, or regional variations. Sentiment analysis applied to textual data could unveil opinions or sentiments expressed, providing insights into user feedback or public sentiment. External environmental factors like weather conditions or pollution levels may be pertinent, significantly if they impact the subject matter. Customer satisfaction survey data, competitor information, and social media activity metrics could enhance the dataset by offering perspectives on satisfaction levels, industry benchmarks, and public discussions. Depending on the dataset's domain, including health metrics or regulatory compliance data would contribute to a more comprehensive and holistic analysis. Carefully integrating these diverse data types aligns with the dataset's objectives and facilitates a more insightful exploration of the phenomena.