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

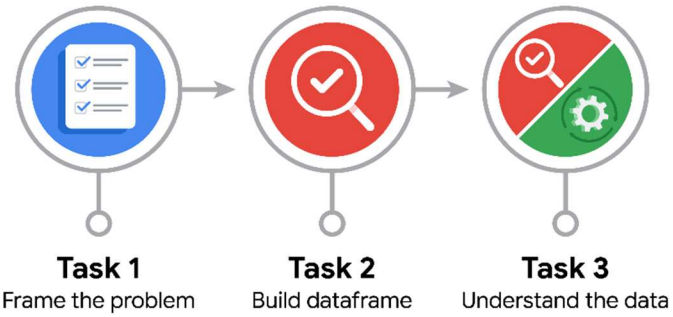
Relevant Interview Questions

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. Ensure that the variables within the dataset provided are understood and any outstanding questions about the data are asked before any analysis begins. 2. Once you've conducted an initial examination of the data, identify ways to structure it for easier analysis. Consider renaming columns, combining or disregarding certain values, and determining the appropriate data types for each column.

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

Examining, exploring and formatting the provided data is essential during this stage. Therefore, any previously created codebooks that cover functions, methods, and tools to summarize, manipulate and format data will be helpful.

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

1. It is important to become acquainted with the dataset and the coding approaches needed to achieve your desired outcomes. 2. To establish a strong foundation for your coding, it is recommended to refer to learning materials or similar work or utilize online resources; a similar problem has likely already been solved.

PACE: Analyze Stage

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

Based on the information provided, it seems like we have everything we need to achieve our goal for this stage. Our objective is to format, restructure, and conduct preliminary analysis of the dataset to uncover useful variables and insights. This will pave the way to a complete exploratory data analysis of the dataset.

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

Utilizing the df.describe() function lets you glean information about the numeric variables in a data frame, such as the minimum, maximum, mean and standard deviation. Furthermore, this provides a window into the distribution of the variables in the dataset, which is important when determining which inference methods to use.

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

The presence of potentially erroneous outliers in the data has moderately skewed the average of certain numeric variables. However, there are a large amount of observations in the data, and therefore the nearly normal condition should be upheld. The interval data is represented by the 1st Quartile, 2nd Quartile, and 3rd Quartile, which indicate where the corresponding data falls within the range of 25%, 50%, and 75% of the 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?

Upon conducting an initial analysis of the dataset, I recommend the following actions to be taken:

Conduct a thorough investigation into the outliers of specific variables, including the minimum and maximum values found in the tip_amount, trip_distance, and fare_amount columns. From a preliminary standpoint, it appears that for the purposes of this analysis, these variables will be important will require further investigation.

- What data initially presents as containing anomalies?

There are discrepancies in the fare_amount, trip_distance, and tip_amount columns that require further investigation before analysis can be completed. It is necessary to confirm if these extreme outliers are errors by the company and should be removed or if they are legitimate.

- What additional types of data could strengthen this dataset?

It would be insightful to have data on the trip times to analyze the most and least number of trips taken by the day. This could also lead to implementing surge pricing during high-peak hours and savings during off-peak hours.