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.

Task 1
Frame the problem

Task 2
Build dataframe

Task 3
Understand the data
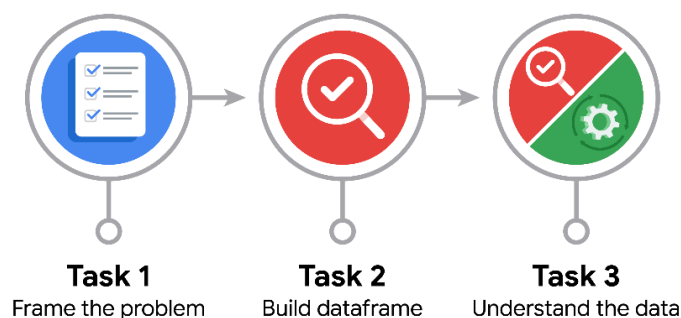
Data Project Questions & Considerations

PACE: Plan Stage

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

  I can best prepare to understand and organize the information by initiating an exploratory data analysis using Python.

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

  The Python codebook in JupyterHub would be a helpful follow-along tool to aid me into performing this work.

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

  Understand the data literally and review the methods of Python code, such as view the data dictionary, reviewing Python methods and the two important libraries that involves dataframes, Numpy and Pandas, before starting to code.

PACE: Analyze Stage

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

> The available information would be sufficient to achieve the goal through simple EDA. But further information is needed when I need to handle more things such as negative values and errors in data imputation.

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

> I built the summary dataframe using the Numpy and Pandas libraries and uses df.describe().

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

> Yes, some of the averages of the data variables look unusual. The max values can significantly and disproportionally exceed the mean and median of some variables, which is highly unusual.

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?

> I would recommend my manager to investigate further into data imputation issues before doing any EDA, because having a negative cost value could be more concerning than just empty or "unknown" values.

- What data initially presents as containing anomalies?

The cost data columns have anomalies because of its negative values.

- What additional types of data could strengthen this dataset?

Additional types of data that could strengthen this dataset could be vehicle types (a categorical type) or simply the cost of the taxi used since different types of vehicles could imply different costs, and could explain why some shorter-distance trips cost more.