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

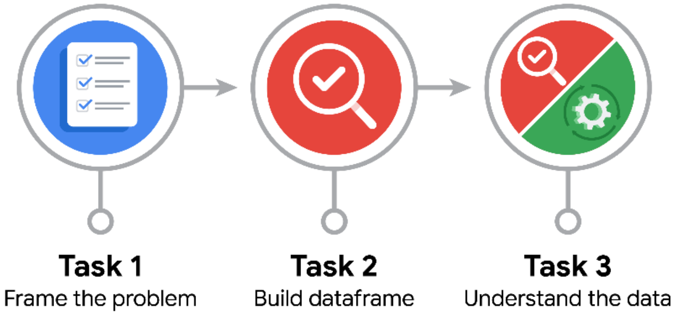
Relevant Interview Questions

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?

The following are some of the steps I would take to prepare to understand and organize the provided taxi cab information:

- Define the problem

- Identify the audience of the information

- Consider the context to which the cab information is to be used. This may involve understanding business objectives and specific requirements and constraints

- Provide concrete examples of how the taxi cab information will be used. This can include specific KPI and specific scenarios of the practical application of the data

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

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

Before diving into code, a resourceful learner would engage in various activities to enhance their understanding, preparation and efficiency. Here are some examples activities that might be performed before diving into code:

- Define clear objectives
- Research and gather related information
- Sketch a plan or workflow
- Check for existing solution
- Review requirements and constraints
- Set realistic milestones
- Consult with teammates
- Allocate time and prioritize tasks

PACE: Analyze Stage

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

Yes, provided we can get a good explanation for some of the anomalies in the data set

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

With a method like describe() in pandas, I can generate summary statistics and assess the min and max range of the data.

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

Yes, for some of the numerical variables the average looks unusual.

For instance, the mean fare amount is 13.03, however the minimum and maximum are -120.0 and 999.9 respectively. This looks unusual upon initial analysis. I observed similar intervals with total amount, tip amount

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 investigating the anomalies related to fare amount, tip amount, and total amount with min and max values far from the mean. For instance, the minimum fare amount is -120 while the maximum is 999.99.

Also, it would be nice to know why “dispute” is coded as a payment type.

Another recommendation would be to ask who the vendors are. I think we can derive some context from learning about this information.

Lastly, I would recommend transforming some of the datatypes. For instance, VendorID should be categorized but it is currently a numeric. The date variable is currently an object, it should be converted to a datetime datatype for easy manipulation.

- What data initially presents as containing anomalies?

The fare amount, total amount contains some negative values. Which are considered to be anomalies as the fare rate can’t be negative.

Also, one trip stands out among the rest with a high total amount (negative).

- What additional types of data could strengthen this dataset?