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

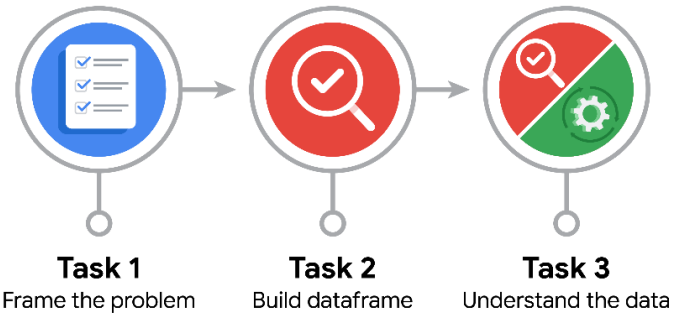
Completing the end-of-course project will help you respond these types of questions that are often asked during the interview process:

- Describe the steps you would take to clean and transform an unstructured data set.
    - Clean up text. Eg. Remove whitespace and special characters, format text
    - Identify, then correct or remove missing or duplicate/missing values
    - Convert values to appropriate data types
    - Create additional calculation columns as needed
- What specific things might you look for as part of your cleaning process?
    - Outliers and anomalies
    - Duplicate or missing values
- 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?

- Outlier/anomaly values that fall outside the logical range of values (eg. A negative value for age, a building height that is 0.8m tall)
- Duplicate or missing values where there should not be any (eg. Duplicate entries of a unique ID)
- These anomalies and outliers have the potential to skew the impact that the variable has on the project goal. For example a negative age will skew the average age of the userbase to a lower value which may cause the analysts to misidentify the target demographic for their product.
- There is an unnamed column at the beginning of the dataset that is not present in the data dictionary. Without knowing what this data represents, we are unable to analyze it and it's relation to the rest of the data.

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?

  Review the data dictionary to understand the fields used in the dataset.

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

  Jupyter Notebooks

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

  Review business question and identify the goals of the project

PACE: Analyze Stage

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

Yes

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

The pandas function `df.describe()` provides a statistical summary of the numerical data in the dataframe, including mean, interquartile ranges, min and max values.

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

The average of cash tip amounts is 0, which is unusual in American society as tipping is a large part of the culture.

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 further analysis of the correlation of total_amount and trip_distance variables as this pertains to what we are investigating (cost of taxi fare).

- What data initially presents as containing anomalies?

Total_amount contains payments of negative amounts, and trip_distance contains taxi rides of 0 distance. There is also an unnamed column present in the dataset that is likely a mislabeled column of Trip ID.