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?

First, I will read the project instructions thoroughly to understand the goals and deliverables.

I will examine the provided data dictionary (if available) or the column headers to understand the meaning of each variable.

I will make notes about the expected data types and potential data quality issues.

I will plan to do an initial inspection of the data using head(), info(), and describe() methods in pandas.

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

Codebooks and notebooks from previous lessons on data cleaning, data transformation, and exploratory data analysis (EDA) will be very helpful.

Specifically, notebooks that demonstrate:

Handling missing values

Data type conversions

Outlier detection and handling

Descriptive statistics and visualizations

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

Research the TLC (Taxi and Limousine Commission) and the context of the data (e.g., time period, location).

Consider potential biases or limitations in the data.

Brainstorm potential questions or hypotheses that can be explored with the data.

Research the specific datatypes that are expected within the data.

PACE: Analyze Stage

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

This depends on the specific goals of the project. Initially, I will assess if the variables provided are relevant to the questions being asked.

I will also look for missing data or inconsistencies that might limit the analysis.

After an initial look at the data, I can better assess if the data is sufficient.

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

I will use the `describe()` method in pandas to generate summary statistics (count, mean, std, min, max, quartiles).

I will examine the min and max values to identify potential outliers or data entry errors.

I will examine the dtypes of each column, to ensure they are the correct type.

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

I will compare the averages to expected values or industry benchmarks (if available).

I will analyze the distribution of interval data using histograms and box plots.

I will look for any data that is outside of a normal range.

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:

Missing values in key variables (e.g., trip distance, fare amount).

Potential outliers in trip duration or fare amount.

The distribution of pickup and dropoff locations.

The distribution of the number of passengers.

I would recommend validating the data against external sources, if possible.

- What data initially presents as containing anomalies?

Zero or negative trip distances or fare amounts.

Extremely long trip durations or unusually high fare amounts.

Inconsistent date or time formats.

Any null values in key columns.

- What additional types of data could strengthen this dataset?

Weather data (to see if weather affects ridership).

Traffic data (to understand trip duration).

Location data for points of interest (to identify popular destinations).

Holiday or event data, to see how special events affect ridership.

Demographic data could provide additional insights.