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. As a reminder, this document is a resource that you can reference in the future and a guide to help consider responses and reflections posed at various points throughout projects.

Course Project Recap

Regardless of which track you have chosen to complete, your goals for this project are:

- ☐ Understand and assess the proposed scenario
- ☐ Demonstrate understanding of the form and function of Python
- ☐ Show how data professionals leverage Python to load, explore, extract, and organize information through custom functions
- ☐ Articulate findings in a professional summary for cross-functional team members

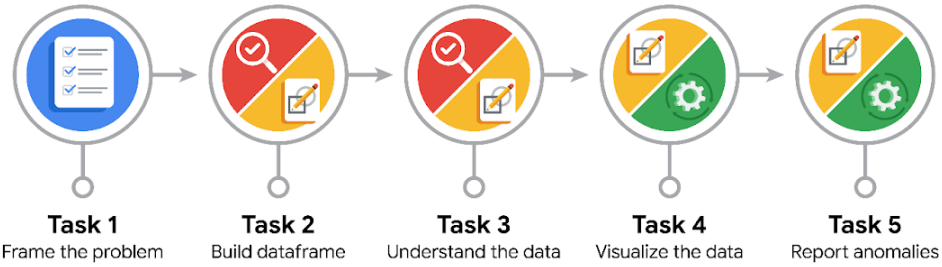
Relevant Interview Questions

Completing this portfolio project will empower you to respond to the following interview topics:

- 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 five tasks; the visual below identifies how the stages of pace are incorporated across those tasks.



Data Project Questions & Considerations

PACE: Planning Stage

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

I have to import and evaluate the data in Python, so I can be able to analyze the data and detect trends, abnormalities, and insights for the analysis.

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

The main codeblocks to perform the initial analysis are destined to import the DataFrame, libraries, and perform general analysis with functions such as head() or info(). I will use Pandas and Numpy to analyze the data and observe the key features such as trends or missing values.

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

Import the libraries and prepare the pseudocode of what he is going to do. The pseudocode will be essential to organize the code in a less consuming and more efficient way.

PACE: Analyzing Stage

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

The information obtained by the general summary of the data frame has helped to understand the nature of the database, however, it is not enough to understand trends and key features of the relationship between variables. Additionally, it seems that the data should be cleaned and organized because there are outliers which do not seem to be right due to the main tendency. Finally, there seems to be a positive correlation between the distance, cost, and zone of the trips.

PACE: Constructing Stage

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

It can be done with the describe() function from the pandas library which shows the main trends and distributions of the data frame.

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

It seems that the means are almost correct. However, there is a real problem with the outliers because there are cases where there is no distance of a trip but there is a fare collected, which may indicate that the distance was not recorded. Additionally, there are trips where the fare was lower than zero, which can be an error at the moment of collecting data. Furthermore, the interval of the data seems to be ranging from 1 to 3 miles of distance per trip and a fare of 8 to 17 dollars per trip if we evaluate the quartiles corresponding to the 25% and 75% and it is corroborated visually by the histograms.

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 will strongly recommend the manager to investigate those missing values and outlier. I would recommend evaluating its causes and justification before cleaning. Additionally, I would recommend evaluating the margin of error of the data.

- What data initially presents as containing anomalies?

Data containing zero values which have the register of a fare but not a distance, and vice versa. Another anomaly is the extreme values which seem to have collecting errors.

- What additional types of data could strengthen this dataset?

Margin of error, the units of the variables, proximity of the zones to the center of the city or recurrent places.