# 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

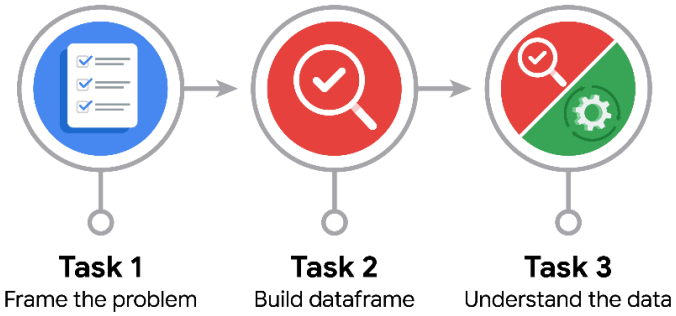
## Relevant Interview Questions

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?

| |
|---|
| To best prepare to understand and organize the provided information, one acquaints oneself with the data. We *begin by exploring our dataset and consider reviewing the Data Dictionary.* In addition, building the necessary dataframe and understanding the data types of its columns is crucial for effective data manipulation and analysis. Once we have a clear picture of our data types, we can delve deeper into understanding and organizing the information. Using describe(), we get summary statistics for numerical columns where we not only analyze mean, median, standard deviation, min, and max values to understand data distribution but identify potential outliers or missing values. |

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

| |
|---|
| Understanding and organizing one's data after examining data types involves a series of steps. Here are some codebook-like structures to guide our process: With respect to Descriptive Statistics and Data Exploration, we import necessary libraries, check data types, descriptive statistics, handle missing values, explore data distribution, and correlation matrix. With respect to Data Visualization and Exploration, we import visualization libraries and create various visualizations. |

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

| |
|---|
| Beyond examining data types, a resourceful data professional would engage in these activities before diving into code: With respect to Deeper Data Exploration, we employ statistical methods or visual techniques to detect outliers. We conduct a thorough check for inconsistencies, missing values, and erroneous data to assess the quality of the data. With respect to Tool and Environment Setup, we set up the necessary development environment and choose appropriate libraries and packages based on the project's requirements. |

## PACE: Analyze Stage

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

| |
|---|
| Although it is possible to determine whether available information is sufficient to achieve the goal during the Analyze phase based on intuition and variable analysis, it is important to approach this assessment with a critical eye and employ rigorous methods. First, since having enough data points is essential for reliable results, the visualizations and summary statistics in the preliminary data analysis phase indicate the amount of data available for analysis is around 19, 000. Second, since the extent of missing data can impact analysis, they indicate the amount of missing data is at most 300 for some variables in the dataset which appears relatively low. However, we need to consider and proceed with exploratory data analysis to reveal further data patterns and potential gaps in the dataset, |

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

| |
|---|
| First, we use Python and Pandas. Second, we utilize the describe() function that provides a comprehensive overview of our numerical data such as count, mean, std, min, max, and percentiles. We assess the min and max range of the data by calling the min() and max() functions that provide the minimum and maximum values for each column, respectively. For categorical data, use functions like value_counts() to understand data distribution. |

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

| |
|---|
| For video_view_count, video_like_count, and video_share_count look unusual because they have large standard deviations. Plus, they have small min values and large max values given their means and medians (i.e., their 50th percentiles). This is indicative of outliers that exist in the data. Plus, we see that although the mean and median for each different claim status are similar to each other there is huge discrepancy between the view counts of claims and opinions |

## 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 initially recommend performing further data cleaning and data analysis steps to understand unusual variables and outliers. We need to decide whether to remove or cap outliers and whether to impute or drop rows with missing values based on their impact on analysis.<br><br>I initially recommend that the data team use descriptive statistics to gain a better grasp of the data and conduct a complete exploratory data analysis. We need to dive deeper into specific variables or subsets of the data. |

- What data initially presents as containing anomalies?

| |
|---|
| Since the describe() method identifies the distributions of each variable such as the standard deviation, percentiles, mean, and median (which is the 50th percentile). It helps as well in indicating any questionable values and outliers by providing the min, max, and standard deviation. For example, it seems that there are outlier values for video_duration_sec since it has a min of 5 when it has a mean and median around 32 and a relatively high standard deviation around 16. Plus, we see that video_view_count, video_like_count, video_share_count, and video_comment_count have relatively high standard deviations given their mean and median and large max values. |

- What additional types of data could strengthen this dataset?

| |
|---|
| To enhance the dataset and model effectiveness, we need to consider Feedback Data such as user ratings, clickstream data, survey responses and Contextual Data such as location data, time-related data, and demographic information. The Contextual Data may indicate nefarious actors such as bots. |