**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 prepare effectively, the following steps need to be taken:

* **Review the dataset structure**: Familiarize yourself with the columns, data types, and context of each variable (e.g., views, likes, shares).
* **Define a clear problem statement**: For example, “Identify what factors most strongly influence TikTok video engagement.”
* **Set goals**: Determine what insights or visualizations you want to generate from the data.
* What follow-along and self-review codebooks will help you perform this work?

Course 2 TikTok Project Lab notebook on Coursera (provided in the lab).

Pandas and NumPy documentation for reference on data manipulation.

Seaborn and Matplotlib documentation for creating visualizations.

Google’s Data Analytics Certificate cheat sheets (if you have access) or Coursera-provided PDF summaries.

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

**Conduct preliminary research** on TikTok engagement metrics to better understand the business context.

**Sketch out a data workflow** on paper or a whiteboard (e.g., import → clean → analyze → visualize).

**Read metadata or data dictionaries** if available, to clarify ambiguous column names.

**Plan hypotheses** to test, like “Do videos with more hashtags receive more shares?”

**PACE: Analyze Stage**

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

Yes, the dataset appears to include key engagement metrics (views, likes, shares, comments) and characteristics of the content (duration, hashtags, description), which are sufficient to draw meaningful insights on engagement.

Start by importing the packages that you will need to load and explore the dataset.

import pandas as pd

import numpy as np

import seaborn as sns

import matplotlib.pyplot as p

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

You can use:

python

Copy

Edit

df.describe()

This command provides summary statistics (mean, std, min, max, 25/50/75 percentiles) for all numeric columns.

To specifically check min and max:

python

Copy

Edit

df.min()

df.max()

Or per column:

python

Copy

Edit

df['views'].min(), df['views'].max()

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

Averages may be skewed if a few videos went viral (millions of views) while most others didn’t, causing a high mean but low median.

Use:

python

Copy

Edit

df['views'].hist(bins=50)

to assess distribution visually.

Interval data includes:

views, likes, shares, comments: Continuous numerical values.

These should be assessed for outliers, distribution skew, and potential normalization.

**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.

N/A

**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?
* **Investigate the distribution** of engagement metrics to identify if certain types of videos (e.g., short vs. long, trending hashtags) consistently outperform others.
* **Verify data cleanliness**: Are there duplicates, missing values, or mislabeled fields?
* **Explore posting patterns**: Are there temporal trends (e.g., time of day, weekday vs. weekend)?
* What data initially presents as containing anomalies?

**Extreme outliers** in views/likes/shares could skew the results (e.g., a few viral videos).

**Missing or null values** in hashtag or description fields.

**Zero or negative values** in engagement metrics (if they exist).

* What additional types of data could strengthen this dataset?
* **Video category or genre** (e.g., comedy, education, dance).
* **Geographic location** of where the video was posted or viewed.
* **Follower count of the creator**—more followers likely equals higher baseline engagement.
* **Posting time and date** as separate columns (if not already present).