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~~

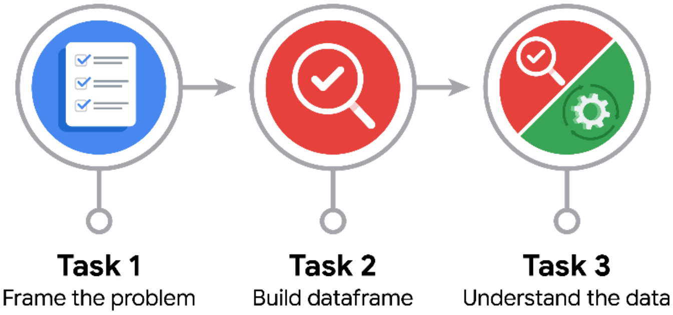
Relevant Interview Questions

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?

Read the information table of each variable from the data set. Look for correct variables type and formatting, check for null values, duplicates, outliers, strange or non-sense values, and reading key words in the transcription that can help finding patterns related to tagging a text as opinion or claim.

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

Especially the last ones regarding grouping and aggregation and boolean masking.

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

Read about TikTok Safety measures again to review verification and banning status in order to have a better understanding of the dataset. Read the emails received again, and checking tasks

PACE: Analyze Stage

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

Is there any tags or topics associated with the transcript text? For instance: politics, beauty, health, ... That would be interesting to have those for getting more detailed and classified insights.

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

I will use methods like, info(), describe(), value_counts()

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

Except by the video duration, the rest of numeric variables have a standard deviation so much greater than the mean values. That shows considerably long ranges for those features. Although that could also be a sign of outliers, it's not clear if there are any yet. Besides, the data doesn't have non-sense values like a negative amount of dowloads or likes, so far. However, there is a strange video of 5 seconds, which is a very short and uncommon period, and it seems it not easy to do a claim or opinion in a five seconds duration. It will be immportant to analyse that video and perhaps some other with very few seconds.

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?

Change the data type of the interactions variables (likes, shares, comments, downloads) to integers.

Look for more information of all the numerical variables, except by video’s duration, since their standard deviation is much more greater than their average values.

Remove the entries that have NaN values. They don’t contain the information we are interested in and represent less than 2%.

Verified accounts are not a representative part of the dataset so this may be discarded as a variable to work with in this analysis. They are just about 6 %.

Columns ‘#’, ‘video_id’ should be removed, since they don’t provide any important information.

Check duration time videos, especially the ones around 5 seconds.

Key words in the transcriptions that can help finding patterns related to tagging a text as opinion or claim.

- What data initially presents as containing anomalies?

The variables that don't contain null values are: #, video_id, video_duration_sec, verified_status and author_ban_status. It seems that in the cases in which there is no transcription text, reasonbly, there is no claim status, but surprisingly there aren't counting of views, shares, likes, downloads or comments.

Except by the video’s duration, the rest of numeric variables have a standard deviation so much greater than the mean values. That shows considerably long ranges for those features. Although that could also be a sign of outliers, it's not clear if there are any yet. Besides, the data doesn't have non-sense values like a negative amount of dowloads or likes, so far.

However, there is a strange video of 5 seconds, which is a very short and uncommon period, and it seems it not easy to do a claim or opinion in a five seconds duration. It will be immportant to analyse that video and perhaps some other with very few seconds.

- What additional types of data could strengthen this dataset?

Percentages of interaction variables (likes, shares, comments, downloads) over view count; topic or tag which classifies a video’s content to check possible correlations with that. General statistics of banned and under review authors. Key words in the transcriptions that can help finding patterns related to tagging a text as opinion or claim.