Course Three

Go Beyond the Numbers: Translate Data into Insights

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 3 PACE strategy document~~
- ☑ ~~Answer the questions in the Jupyter notebook project file~~
- ☑ ~~Clean your data, perform exploratory data analysis (EDA)~~
- ☑ ~~Create data visualizations~~
- ☑ ~~Create an executive summary to share your results~~

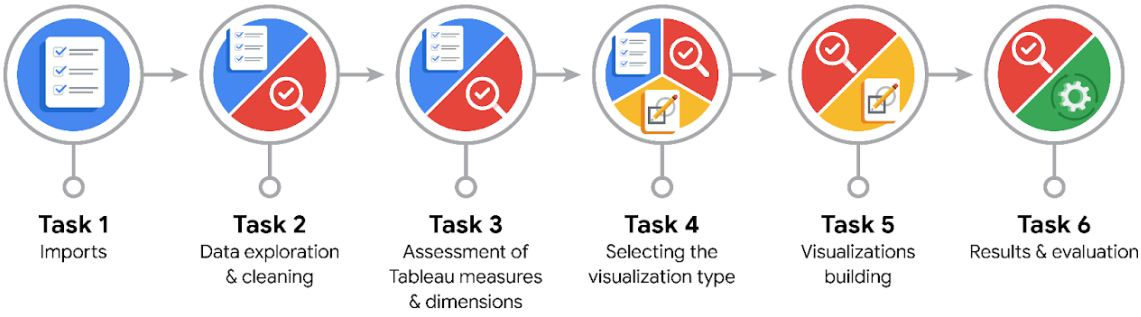
Relevant Interview Questions

Completing the end-of-course project will help you respond to these types of questions that are often asked during the interview process:

- How would you explain the difference between qualitative and quantitative data sources?
- Describe the difference between structured and unstructured data.
- Why is it important to do exploratory data analysis?
- How would you perform EDA on a given dataset?
- How do you create or alter a visualization based on different audiences?
- How do you avoid bias and ensure accessibility in a data visualization?
- How does data visualization inform your EDA?

Reference Guide

This project has six tasks; the visual below identifies how the stages of PACE are incorporated across those tasks.



Data Project Questions & Considerations

PACE: Plan Stage

- What are the data columns and variables and which ones are most relevant to your deliverable?

There are 12 columns in the provided dataset, most of which include the metrics used to measure the viewers engagement on a video. These metrics are: number of views, likes, comments, shares, and downloads. In addition to these columns we also have the categorization of each video as a claim or opinion, where each status is either verified, or not verified. Finally the last column that is of interest is the 'author ban status', because this will help us give some context as to what is happening with these two categories and how more often than not claim videos have banned authors.

- What units are your variables in?

The numeric units of my dataset are counts which are dimensionless. The other pieces of data are categorical variables, which means they depend on the context of the analysis in how they are utilized.

- What are your initial presumptions about the data that can inform your EDA, knowing you will need to confirm or deny with your future findings?

Some initial presumptions about this data could be that all categorization has been done correctly and that there is no mislabeling with the “claim status” nor “author ban status” columns. Another presumption is that all videos have been categorized into claim or opinion and that there are not entries with Null values in this column.

- Is there any missing or incomplete data?

There wasn’t any missing data, all columns did not have any Null values in them, as informed by the pandas method .info()

- Are all pieces of this dataset in the same format?

Not all pieces of this data set are in the same format, some are numeric and some are objects, which are just strings used to categorize the videos, such as ‘claim status’, ‘verified status’ and ‘author ban status’.

- Which EDA practices will be required to begin this project?

Examine the distribution of the engagement metrics and segregate these metrics based on the videos claim status. Another important aspect we would need to explore are the outliers present in these metrics and ways to manage them. We could use box plots to visualize the number of outliers per metric and define a threshold for us to cut outliers from the analysis.

PACE: Analyze Stage

- What steps need to be taken to perform EDA in the most effective way to achieve the project goal?

To perform EDA in the most effective manner we would ensure that we are accounting for all the engagement metrics provided to us and how they represent the different video categories.

Additionally, we would need to verify that our data is cleaned from outliers or data points that can possibly introduce a bias.

- Do you need to add more data using the EDA practice of joining? What type of structuring needs to be done to this dataset, such as filtering, sorting, etc.?

I don't believe we need additional data based on the exploratory data analysis practices performed on the current dataset. However, it would be more beneficial to sort the data on the claim status as this will help read and compare various metrics for these video categories.

- What initial assumptions do you have about the types of visualizations that might best be suited for the intended audience?

Bar graphs are good ways to represent comparisons between variables, so we should use them when looking at the metric for each video's different claim status. Another useful visualization for comparison of variables are pie charts and box plots, with the latter one providing more detailed information about the statistics of the data.

PACE: Construct Stage

- What data visualizations, machine learning algorithms, or other data outputs will need to be built in order to complete the project goals?

As formerly mentioned, bar graphs, box plots and pie charts will be useful to compare the engagement metrics for the two classes of videos in Tik-Tok. We will also need to implement a filter to remove outliers from the dataset such that the visualizations are not biased.

- What processes need to be performed in order to build the necessary data visualizations?

It's always good to structure a new dataframe with the groupings and columns desired for the visualization. Most of these data frames will just be created using the .groupby() method and then

calling the .agg() method to apply a specific action on the grouped data, such as summing, or counting or averaging.

- Which variables are most applicable for the visualizations in this data project?

The most important variables for the visualization are the numeric variables quantifying the viewer engagement levels per video class.

- Going back to the Plan stage, how do you plan to deal with the missing data (if any)?

If there was any missing data then I would try to remove it or find a form interpolation to fill in the value with an appropriate estimate.

PACE: Execute Stage

- What key insights emerged from your EDA and visualizations(s)?

All the metrics for quantifying the engagement levels are heavily right-skewed. This could be an indication for removing outliers. To improve the visibility of the bar graphs I set the x-limit to the 99-percentile as a relaxed filter for outliers; however, this provided a superficial solution. It would be better altogether to remove the outliers in order to focus primarily on the main trends of the data.

I also found that videos that are categorized as 'claim' have much higher engagement rates than those classified as 'opinion'. However, the ratio of banned authors to active authors is much higher for videos classified as 'claims' than for other videos. This would suggest to implement higher caution with 'claim' authors given this tendency.

We also see that banned authors have much higher engagement metrics than active authors which could be an indication to the type of content that goes viral on the platform.

We also saw that the comment to view relationship for both claim videos and opinion videos follows a linear trend where most of the data points are below the linear trend.

- What business and/or organizational recommendations do you propose based on the visualization(s) built?

I would recommend adding a filter for 'claim' videos to add another step of security regarding the content these types of videos produce on the website, because we see a high correlation between 'claim' videos and 'banned authors'. It's clear these types of videos are much more successful than opinion videos however, they encapsulate a bigger risk for the platform.

- Given what you know about the data and the visualizations you were using, what other questions could you research for the team?

I would like to focus on the metric solely for active authors to see how the platform is performing with its current content creators and spot any sort of trend that could shed some light on why so many 'claim' authors are being banned from the platform.

- How might you share these visualizations with different audiences?

I would make sure my color palettes are accessible for anybody with a color blindness disability, and I would try to make my plots simple enough to avoid confusion but I would annotate them to highlight the important aspects of the data.