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~~

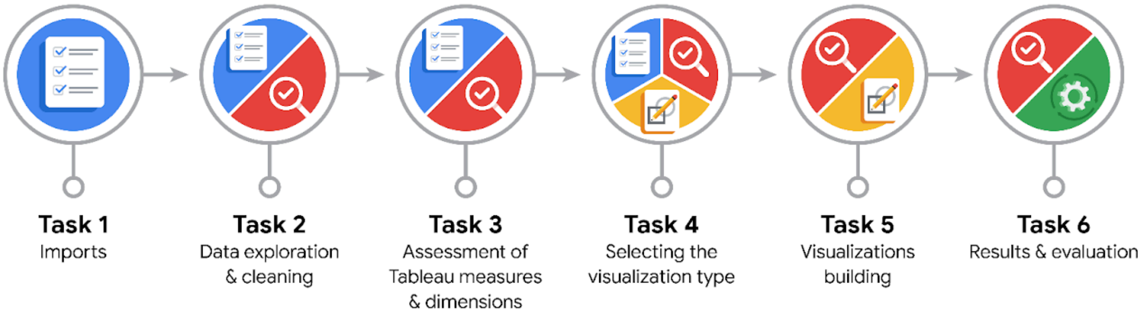
Relevant Interview Questions

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.



Task 1
Imports

Task 2
Data exploration & cleaning

Task 3
Assessment of Tableau measures & dimensions

Task 4
Selecting the visualization type

Task 5
Visualizations building

Task 6
Results & evaluation

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 columns 12 columns and 19.382 videos. Except by ‘video_id’, ‘#’, ‘video_transcription_text’ and, perhaps, ‘verified_status’, the other columns are essential for our EDA, Tableau and executive summary: ‘claim_status’, ‘video_duration_sec’, ‘author_ban_status’, ‘video_view_count’, ‘video_like_count’, ‘video_share_count’, ‘video_download_count’, ‘video_comment_count’.

- What units are your variables in?

‘Video_duration_sec’ is in seconds up to 60 and it is in correct format. The other variables counting videos interaction like views, likes, comments, etc. are integers, however their type is ‘float’, so those will be changed to ‘int’.

- 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?

- There are videos of 5 seconds which could be very short for giving an opinion or claiming something.
- Claiming videos have far more user interactions than opinion ones

- The study must be divided into opinion and claim videos since the latter have a completely our of range interaction counts. Otherwise result will be more skewed to the right.
- There is strong correlation between engagement levels and claim status.

- Is there any missing or incomplete data?

There are 298 videos with null values and they all don't contain any of the most important informations for our purpose like claim status and view, like, share, download and comment count. They don't have any more common characteristics, as was confirmed checking verification and author ban status. The null entries of all those rows are in the same columns.

Fortunately they only represent 1.5 % of the videos in this data set, so removing them it won't affect the sample.

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

The format of each variable is right, though types of some columns need ot be changes as explained before.

Categorical values in claim, verified and author ban status coincide with their descriptions in the Data Dictionary.

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

- Change of type for some columns
- Removing rows with null values since they don't contain the needed information for this analysis.
- Look for abnormal values and outliers
- Nominal categorical variables have at most 3 values, so a one-hot encoding will work for traning a model later.

PACE: Analyze Stage

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

- Delete rows that don't contain the essential information for this project.
- Understand the distribution of the variables.
- Look for correlations between variables, especially related to claim status.
- Identify outliers and decide how to treat them.

- 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 think I need more data for now. The data set must be filtered by claim status when presenting data. Also, using other categorica

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

- Box plots and histograms for understanding engagement variables and video's duration.
- Scatter plots for relationships between two engagement variables.
- Pie charts for percentage of videos in specific categories
- Heat maps and stacked bar charts when having two categorical variables and one numerical variable, like claim status, author ban status and number of video's views.

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?

- A pie chart comparing number of videos by claim status.
- Boxplots for each count variable and for video's duration
- Scatterplot for number of video likes vs views
- Heatmap for comparing number of views by claim status and author ban status together.
- Stacked bar chart for number of videos classified by author ban status and claim status together.

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

- Consider to grap opinion and claim videos apart, since the latter have a very wide range with respect to the former.

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

All the count variables and the nominal categorical ones: claim, author and verified status.

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

I'm going to remove it since it doesn't contain a single value of the important variables.

PACE: Execute Stage

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

- Duration does not have a clear correlation with any of the other variables.
- Claim and opinion videos completely differ in engament's levels which will be helpful when building our predictive model. Claim videos have very high interaction levels.
- Counting variables are correlated with the number of views, as expected.
- Number of views highly correlate with authors status, where active ones have very few views compared to banned and under review status.
- In this case I should keep outliers since they haven't come from misspelings or mistakes, but the behaviour of social networks engaments.
- There is higher probability a verified account posts an opinion than a claim.
- A verified status is correlated with active authors. In banned and under review accounts there are very few verified status.

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

- Exclude video's duration since it doesn't seem as a relevant variable.
- Work deeply in relationships among variables but separating the data set by claim status.

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

- How do we involve the transcription text on this? Some keywords that classify videos as claim or opinions.
- Having decided what values are outliers on each variable and, it seems, knowing we will need ot keep them, how are we going to use that information when traning the model?
- Those outliers represent between 12 to 20 % of the variables, what to do with that?
- Is there something else I need? Like a joining more data?
- Inquiring for more explicit relationships among numerical variables by using a correlation heat map.
- How many of the videos have been classified wrongly? Is there any way to know that information?

- How might you share these visualizations with different audiences?

- With the more technical group I will use the visualisations and analysis I did in Python's lab, whilst I will share the Tableau visuals with the nontechnical members.
- Keep in mind colours that can be easily distinguish for people having visual impairments issues
- Few information on each Tableau Story, and writing conclusions that are no explicitly said by graphs like, instead of an obvious idea, I could give more details related to the business purpose.