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

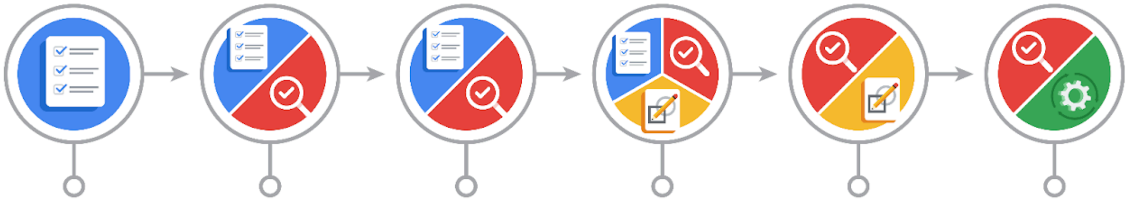
Relevant Interview Questions

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?

Video_view_ count, and all other consumer metrics, then the claim_status, author_ban_status, and verified_status

- What units are your variables in?

The units of all consumer metrics are integers and float, the qualitative variables are categorical strings.

- 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?

That there will be numeric differences between videos that are a ‘claim’ vs ones that are a “opinion”. This will ensure that we will find a way to train a ML algorithm to identify these categories eventually.

- Is there any missing or incomplete data?

There are ~200 hundred rows that contain null values. They are listed with a null 'claim_status' which likely points to them being outside our scope of necessity

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

No, there is int64, float64, and object

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

Assess null values, check for spelling errors in the object columns, assess outliers

PACE: Analyze Stage

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

We need to inspect the data and build visualizations to start bringing the story of the data into the perspective.

- 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.?

No additional data is need for this set. Filtering and sorting the set to see the data before building visualizations can bring a further context to the graphics.

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

We understand that there is not time data, so no chronological time series graphics are needed. There is question of relationship between data, so potentially scatter plots, bar plots, and histograms. Additionally, with outliers being an issue, box plots should be used to better understand the data's distribution.

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?

We must build data visualizations that will display relationships between variables to assist determining the difference between an opinion or claim. Once we find a subset of variables to work with we will build a ML model to discern between a submission for opinion/claim metric.

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

Data exploration and cleaning, discussion and inspection of the outliers, elimination of null values.

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

Most applicable variables are views, likes, comments, downloads show distinct dichotomy between opinion and claim. Additionally, author ban status and author verification show correlation.

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

The missing data could potentially be fixed through checking our data collection logs/system. Why were these 200 values missing when most of this data seems like it is automatically collected, without much human interaction.

PACE: Execute Stage

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

That opinion submissions had much lower engagement, while claims were very viral, banned accounts were more likely to have claim submissions, and verified accounts were more likely to have opinions.

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

I recommend that we move forward with ML modeling, since there seems to be multiple variables that can contribute to correctly categorizing opinion and claim. Additionally, there are factors controlled by the author(verification and being active/banned) and factors from the public (all the engagement metrics). This many options for designing a ML could make for a good decision tree model to effectively use these variables options in each analysis

- 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 dig deeper into reasons why there is so little engagement with opinion posts, how can we better bin the engagement data, so it is not so right skew/distributed, are there outliers to ignore, are some submissions with zeros in the engagement variable outliers?

- How might you share these visualizations with different audiences?

scatter plots should come with explanation for all audiences since they are hiding the opinion data points, box plots should be offered with extra explanation to non technical staff due to the complexity of outliers, box plots and histograms are somewhat self explanatory but stacking boxplots can need some extra thought.