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

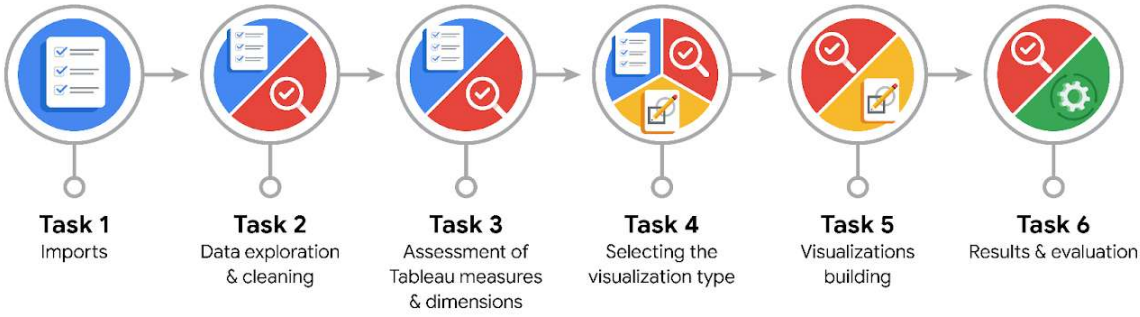
Relevant Interview Questions

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?

The most relevant variables for predicting whether a video contains a claim or an opinion are the dependent variable claim_status and the independent variables author_ban_status, video_view_count, video_like_count, video_share_count, video_download_count, and video_comment_count.

- What units are your variables in?

The variables are of integer (int64), floating-point (float64), and object (string) types.

- 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?

A preliminary investigation of the data reveals a potentially serious issue: banned content appears to generate significantly more engagement than non-banned content. This is especially problematic if misinformation is reaching a larger audience, as it can have detrimental real-world impacts. Moreover, initial analysis suggests a correlation between author ban status and the presence of claims in videos. Specifically, videos classified as containing claims are disproportionately linked to banned authors and unverified users, raising concerns about the source and spread of potentially harmful information.

- Is there any missing or incomplete data?

The dataset contains 298 missing values distributed across seven columns. Further analysis will be conducted to determine the nature of these missing values, specifically whether they are missing completely at random (MCAR), missing at random (MAR), or not missing at random (NMAR).

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

The dataset contains variables of types int64 (integers), float64 (floating-point numbers), and object (strings). The formatting within each of these types appears consistent.

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

The investigation of the dataset will involve calculating descriptive statistics (e.g., mean, median, standard deviation), visualizing variable distributions (e.g., histograms, box plots), and implementing appropriate strategies for handling missing data and outliers, if present.

PACE: Analyze Stage

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

While the order of exploratory data analysis (EDA) steps is flexible, a structured approach is recommended. This involves systematically addressing key EDA practices during data exploration. Thorough initial EDA is crucial, as incomplete exploration can significantly compromise the validity of subsequent statistical tests and model predictions.

- 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.?

At this stage of model development, it is unclear whether augmenting the dataset with additional data will be beneficial. This question will be re-evaluated during a second iteration of model development, after initial model performance has been assessed.

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

Visualizations such as bar plots, pie charts, and scatter plots will be employed to effectively illustrate the relationship between claim status and author ban status. Additionally, visualizations will be used to depict the distributions of count variables, including views, shares, comments, and likes, providing a clear overview for the audience.

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?

For visualization, we will utilize a variety of plots, including pie charts, box plots, scatter plots, and bar plots. The modeling phase will involve testing several algorithms, including logistic regression, decision trees, random forests, and XGBoost.

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

Before developing visualizations, a preliminary investigation of the dataset, including initial exploratory data analysis (EDA), will be performed. Addressing outliers and missing values is crucial, as their presence can skew initial visualizations and misrepresent the underlying data.

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

The distributions and potential outliers of the count variables will be visualized using box plots and histograms. Furthermore, pie charts, bar plots, and scatter plots will be employed to compare different groups within the claim status and author ban status variables, revealing potential relationships and differences.

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

A preliminary analysis of the dataset identified 298 missing values distributed across seven columns. Further investigation is required to determine the missing data mechanism, specifically whether these values are missing completely at random (MCAR), missing at random (MAR), or not missing at random (NMAR). If the missing data is deemed non-problematic and constitutes a small fraction of the dataset, the trade-offs between retaining the affected rows and removing them will be evaluated.

PACE: Execute Stage

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

The count variables exhibit a substantial positive (right) skew, necessitating appropriate transformations. Furthermore, the distribution of claims versus opinions across these variables (likes, shares, views, and comments) is significantly imbalanced, indicating a disproportionate representation of one category over the other.

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

To ensure the validity of subsequent statistical tests and model development, the skew and imbalance present in these variables must be addressed. Additionally, appropriate strategies for handling the missing values will be implemented.

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

We will analyze the video_transcription_text variable to identify trends distinguishing claims from opinions, focusing on keyword analysis. We will also investigate the relationship between video_duration_sec and claim_status.

1. What keywords in video_transcription_text differentiate claims from opinions?
2. Is there a correlation between video length (video_duration_sec) and claim status?

- How might you share these visualizations with different audiences?

For technical audiences, granular visualizations such as box plots will be provided. For non-technical stakeholders, visualizations highlighting overall trends, patterns, and group counts, such as bar plots, pie charts, and scatter plots, will be used.