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

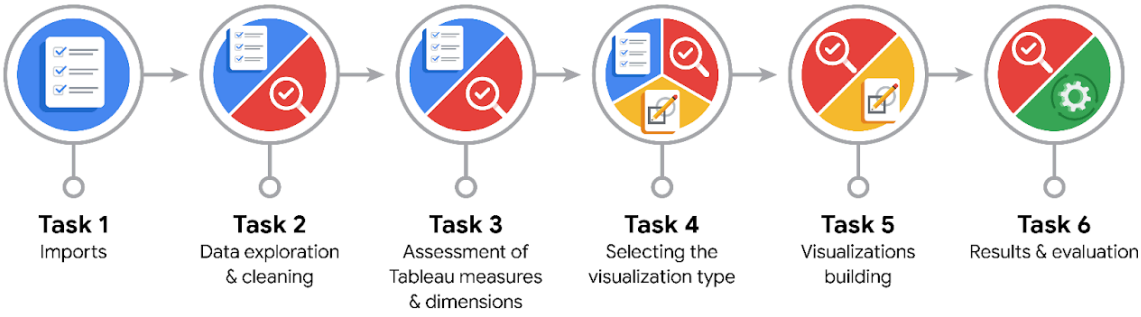
Relevant Interview Questions

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?

Identifying crucial variables for the TikTok Claims Classification project involves focusing on claim and opinion counts and video-related metrics such as duration, likes, comments, views, and the author's ban status. These key variables enable a comprehensive exploration of user-generated content, engagement patterns, and potential correlations with claims and opinions. Selecting and analyzing these variables aligns with the project's EDA objectives and lays the foundation for insightful data visualizations in both Python and Tableau.

- What units are your variables in?

The specific units of the variables in the TikTok Claims Classification project should be explicitly mentioned in the provided information. It would be crucial to inspect the dataset to determine the measurement units for each variable, whether they represent counts, durations, percentages, or other relevant units. This information is essential for accurate interpretation and visualization during the exploratory data analysis (EDA) process.

- 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?

Without specific details about the TikTok Claims Classification dataset, initial presumptions could include anticipating a variety of video-related metrics, such as video duration, like counts, comment counts, share counts, and download counts. Expectations include observing patterns between claims and opinions, potential outliers in critical variables, and insights into the impact of variables like video view counts on the nature of the content. However, these presumptions need verification through the actual EDA process to draw accurate conclusions about the data's characteristics and relationships.

- Is there any missing or incomplete data?

Check for missing values using functions like `isnull()` in Python. Analyze data summary statistics for unexpected patterns or outliers. Visualize the dataset for irregularities. Leverage domain knowledge to understand if missing data is expected. Communicate with stakeholders to address missing or incomplete data, considering imputation or removal strategies.

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

Perform data validation checks to ensure consistency in formats across variables. Examine different data types and formats within each column. Utilize functions like `dtypes` in Python to inspect variable types. Standardize formats for uniformity if discrepancies are found. Document any transformations made for transparency in analysis.

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

To initiate the project, essential EDA practices include: 1. Exploring the dataset's structure and identifying relevant variables. 2. Checking for missing or incomplete data to address potential gaps. 3. Assessing variable types and ensuring consistency in formats. 4. Investigating distributions and patterns in key variables.

5. Formulating initial hypotheses to guide further explorations and confirmations.

PACE: Analyze Stage

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

To conduct effective EDA for the project goal, follow these steps:

1. Develop a clear understanding of the dataset's structure and variables.

2. Address missing or incomplete data through imputation or removal.

3. Ensure consistency in variable formats for accurate analysis.

4. Explore key variables through statistical summaries and visualizations.

5. Investigate relationships between variables to identify patterns and insights.

6. Utilize Tableau for visualizations, focusing on claim/opinion counts and variable comparisons.

7. Incorporate seaborn visualizations in a Python notebook to showcase the cleaning and structuring process.

8. Create a Tableau dashboard for claims versus opinions and stacked bar charts to aid non-technical stakeholders.

9. Provide an executive summary to communicate key findings to the management team.

- 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, joining additional data is not mentioned in the scenario. However, structuring tasks such as filtering, sorting, and creating subsets are needed to enhance the dataset for practical analysis.

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

The intended audience, including management and the assistant director with visual impairments, suggests a need for visually accessible and easy-to-understand visualizations. Bar charts, stacked bar charts, and Tableau dashboards could effectively convey information clearly to a non-technical 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?

The project goals involve performing exploratory data analysis (EDA) and creating visualizations. Relevant visualizations include a graph comparing claim counts to opinion counts, boxplots of essential variables, a Tableau dashboard showing claims versus opinions count, and stacked bar charts for variables like video view counts, video like counts, video share counts, and video download counts. As mentioned in the project scenario, the data scientist should also consider incorporating machine learning algorithms for predictive modeling.

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

To build the necessary data visualizations for the project, the focus will be on cleaning and structuring the data, performing exploratory data analysis (EDA), selecting relevant variables based on insights from EDA, creating visualizations using tools like Matplotlib, Seaborn, and Tableau, and integrating machine learning if applicable for predictive modeling of claims classification. These steps ensure the practical construction of visual representations that address the project's goals and requirements.

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

The most applicable variables for the visualizations in this data project would likely include "claim counts," "opinion counts," "video duration," "video like count," "video comment count," "video view count," "video share counts," "video download counts," and "author ban status counts." These variables

are crucial for understanding the distribution, relationships, and potential outliers in the data, aligning with the project's goals of claims classification and efficient reporting to stakeholders.

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

  To deal with missing data, I plan to employ strategies such as imputation, where feasible, to replace missing values with reasonable estimates. Additionally, I will assess the impact of missing data on the analysis and consider excluding or addressing it appropriately. This process will involve using functions like pandas' `isna()` and `fillna ()` to identify and handle missing values effectively, ensuring the robustness of the exploratory data analysis and visualizations.

PACE: Execute Stage

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

During the Execution stage, critical insights from the exploratory data analysis (EDA) and visualizations revealed patterns in user-generated content on TikTok. The comparison of claim counts to opinion counts illustrated the distribution of content types, with a notable focus on understanding user claims. Boxplots of variables like video duration, likes, comments, and views helped identify outliers and understand the data's variability. Tableau visualizations, including a claims versus opinions dashboard, showcased the distribution across key variables like video view counts, likes, shares, and downloads. These insights contribute to a comprehensive understanding of the platform's user interactions and content trends.

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

Key insights have emerged upon conducting exploratory data analysis (EDA) and creating visualizations for TikTok's claims classification project. The analysis revealed the need to prioritize content moderation for videos with high claim counts, enhancing the platform's efficiency in addressing user reports. Additionally, a focus on boosting user engagement is recommended,

tailoring features to promote content with high metrics like views, likes, shares, and downloads. Managing outliers in metrics such as video duration, likes, comments, and views is essential to maintain a balanced user experience. Considering accessibility features in visualizations ensures inclusivity for users with visual impairments. Lastly, establishing a framework for continuous monitoring and periodic updates will contribute to a proactive approach to addressing evolving content trends and user interactions.

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

The key insights from the TikTok claim classification project's EDA and visualizations reveal patterns in content reporting and user behavior. The project highlights the need for targeted moderation strategies based on content features and regional variations. Business recommendations include optimizing moderation workflows and understanding the impact of content moderation on user engagement. The findings underscore the importance of a nuanced approach to content management for a platform like TikTok. Further research could explore temporal trends, user demographics, and the effectiveness of moderation practices.

- How might you share these visualizations with different audiences?

A strategic approach is crucial to share the visualizations with diverse audiences effectively. For the executive team, concise and visually appealing Tableau dashboards highlighting key performance indicators (KPIs) can aid in quick decision-making. The data science team may benefit from detailed Python notebooks showcasing the EDA process, code snippets, and raw data access for thorough analysis and collaboration. General stakeholders, including those with visual impairments, can engage with interactive Tableau dashboards featuring explicit annotations and tooltips for user-friendly exploration of critical trends and insights. Ensuring compliance with accessibility standards, such as providing alternative text and downloadable data, is essential for inclusivity. Conducting training sessions with documentation and guides fosters a self-service exploration culture, empowering teams to extract valuable insights from the visualizations.