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

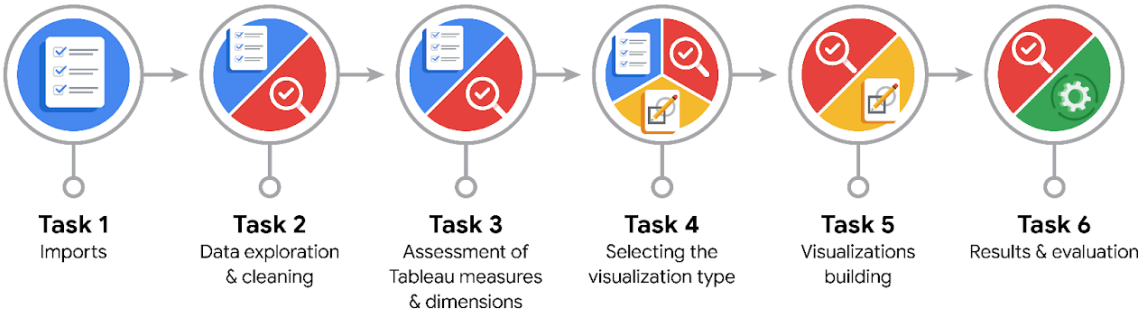
Relevant Interview Questions

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 data columns and variables and which ones are most relevant to your deliverable are: claim_status, verified_status, author_ban_status, video_duration_sec, video_view_count, video_like_count, video_share_count, video_comment_count, and video_download_count.

- What units are your variables in?

These variables do not have units in the traditional sense. Only video_duration_sec has units in the traditional sense. Its units are in seconds.

- 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?

From the previous milestone, the initial presumption about the data that can inform my EDA is that there are roughly the same number of claims and opinions video in the dataset.

- Is there any missing or incomplete data?

Yes, there is missing or incomplete data. Since there are 19382 total rows, there are several key variables (like claim_status, video_view_count, video_like_count, video_share_count, video_download_count, and video_comment_count) with 19084 rows.

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

No, some pieces of this dataset are in different formats (float64(5), int64(3), object(4)).

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

While the core of EDA happens in the Analyze stage, some preliminary EDA can be beneficial in the Plan stage to refine project goals and scope. Here are some key practices: - Via Data Exploration, we create a preliminary data dictionary to define variables and their meanings, grasp a basic understanding of the data's structure, size, and format. We identify potential issues like missing values, outliers, or inconsistencies. - Via Data Visualization, we create basic visualizations (histograms, box plots, scatter plots) to get a sense of data distribution. Thus, EDA helps a data professional to get to know the data, visualize it, understand its outliers, clean its missing values, and prepare it for future modeling.

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 way to achieve the project goal, we need to use functions such as head(), describe(), and info() to analyze the data and matplotlib/seaborn functions/methods to generate visualizations for our end-of-course project.

- 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, it appears that it is not necessary to add more data using the EDA practice of joining. The type of structuring that needs to be done to this dataset are filtering and grouping.

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

The following types of visualizations that might best be suited for the intended audience: Bar charts, Box plots, Histograms, and Scatter plots.

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?

To complete the project goals, bar charts, box plots, histograms, and scatter plots were generated to examine key variables.

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

Applicable packages and libraries were imported to the code notebook such as seaborn and matplotlib.pyplot.

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

The variables that are most applicable for the visualizations in this data project are those most relevant to our deliverables such as claim_status, verified_status, author_ban_status, video_duration_sec, video_view_count, video_like_count, video_share_count, video_comment_count, and video_download_count.

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

Anticipating and planning for missing data is crucial in the Plan and Construct stages of the PACE workflow. During Data Acquisition, we assess the impact of missing data on project goals. And during Exploratory Data Analysis, we plan to deal with the missing data (if any) by cleaning and structuring our data (implementing chosen missing data handling techniques and evaluating the impact of different methods on data distribution and model performance.). Plus, if necessary we may implement common Missing Data Handling Techniques such as Deletion (removing rows or columns with missing values) or Imputation (replacing missing values with estimated values like mean, median, or mode).

PACE: Execute Stage

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

- For many boxplots, boxes of the interquartile range are squished at one end because the outliers at the other end take up all the space. - There appears to be positively skewed distributions for most of the variables - There are far fewer verified users than unverified users, but if a user *is* verified, they are much more likely to post opinions. - For both claims and opinions, there are many more active authors or authors under review than banned authors; however, the proportion of active authors is greater for opinion videos than for claim videos. - The median view counts for banned non-active authors are far greater than the median view count for active authors. It appears that `video_view_count` might be a good indicator of claim status because videos by banned non-active authors get far more views on aggregate than videos by active authors. - The overall view count is dominated by claim videos even though there are roughly the same number of each video in the dataset as indicated in the previous milestone.

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

We need to effectively prepare our data for analysis and modeling, understanding the nature of the outliers and their potential impact and building models with and without them to compare model performance.

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

Some of the questions or concerns one could research for the team is to further investigate distinctive characteristics that apply only to claims or only to opinions; also, other variables that might be helpful in better understanding the data.

- How might you share these visualizations with different audiences?

To share visualizations with different audiences, we not only plot matplotlib/seaborn visualizations to help us understand the data but use Tableau to create visuals for an executive summary to help non-technical stakeholders engage and interact with the data as well.