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

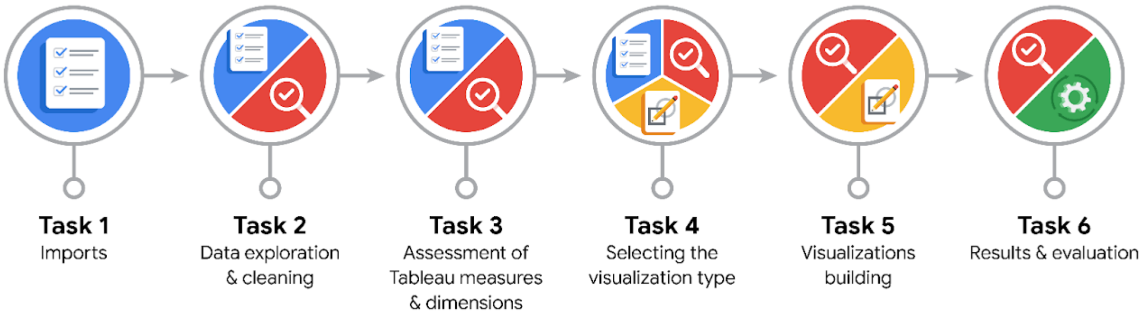
Relevant Interview Questions

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?

Except for these columns: ID, total_navigations_fav1, and total_navigations_fav2, the rest will help work on this scenario.

- What units are your variables in?

Int, float, and object.

- 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?

Bias is one thing to eliminate to produce a high-quality dataset.

- Is there any missing or incomplete data?

Yes, there is missing data, in the label column.

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

No, they aren't. Some show as numbers, and others as strings.

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

Discovering is the first starting point to familiarize oneself with the whole dataset. Use methods and attributes such as head, describe, info, shape, size, etc. to start.

PACE: Analyze Stage

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

Using six practices of the EDA process: discovering, structuring, cleaning, joining, validating, and executing to perform EDA most effectively. These steps don't have to be in the right order and there can be an iterative process that can perform multiple times.

- 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.?

There's no need to add more data. Structuring methods I can use such as slicing, and filtering

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

Barplot, scatterplot, and histogram that these visualizations might be my assumptions to suit the intended 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?

Creating a series of box plots and histograms to visualize.

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

The majority of this project is to build box plots and histograms to visualize. We can choose either maplotlib or seaborn library for visualization. The variables are the main argument we want to visualize each of them. Also, based on the conditions of the arguments we want to build visualizations to fit our needs. Besides, adjusting the plot to clearly view, adding the title and labels for clarification, colors, transparency, etc. to finalize the visualizations.

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

Sessions, drives, total session, n_days_after_onboarding, driving_km_drives, duration_minutes_drives, activity_days, driving_days, and device are the variables

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

It all depends on the amount of missing data. If there's not too much, we might delete them, or we might choose to fill the values (mean or median for numeric, or Nan category for categorical). In case there's a large number of missing, there's no point in continuing with the analysis, and we might contact the data's owner to fill in the values.

PACE: Execute Stage

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

In visualizations, mostly the distributions are right-skew, which is values on the left side the most. Also, there are evenly distributed values.

Users are most likely to churn if they don't use the app and drive at all last month. The opposite is that they tend not to churn when they use the app and drive a lot.

There are extreme values when performing EDA. For instance, the maximum total km a user drives in a day is more than 20,00km, and the maximum value of total duration is ~4,700 minutes/day. These must be investigated further to double-check if something wrong or an error in the dataset.

There are quite a large number of longer tenures that drove a lot last month, which might ask the Waze data team to analyze further.

There's a problem referring to active and driving days. In one of the visualizations, it shows that few users don't open the app at all in a month but there are quite a lot who don't drive at all.

Activity days and driving days, these two variables that are correlated with each other about the churn users rate.

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

Considering and further analyzing the factors that make users choose to churn, especially those who are long-distance drivers.

Research deeper about the users since they tend to churn with fewer or no active and driving days.

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

What factors are contributing to the increased number of users that used the app and drive last month?

What factors make users tend to churn about the fewer or none at all days using the app and driving?

What reasons make users who drive long-distance in driving-days tend to churn?

- How might you share these visualizations with different audiences?

If we want to share with other analysts or managers, we can present fully and explicitly details of what we've done in the visualizations created (for instance we can say most of the visualizations that are right-skew distributions, as well as uniform distribution) with more technical details.

On the other hand, with audiences who are less or non-technical such as executive or cross-functional teams, we want to be the simplest in our presenting of key insights and important findings to them. Avoid getting into more technical details, complex visualizations, etc.