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

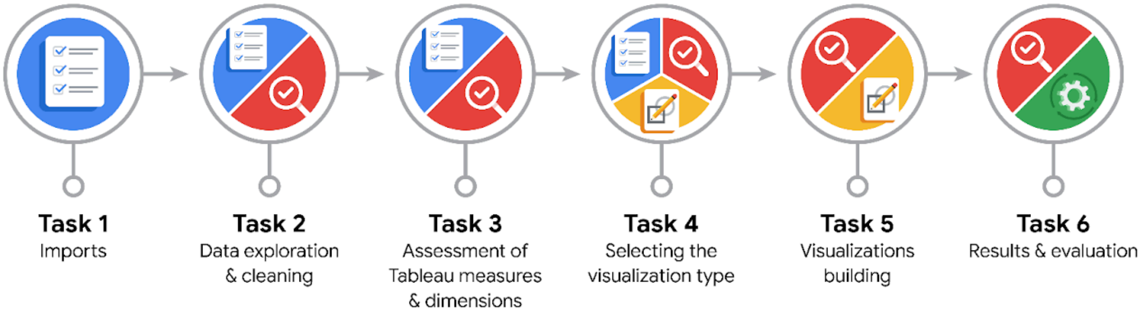
Relevant Interview Questions

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?

Trip duration, total amount, and duration

- What units are your variables in?

Datetiem and float64

- 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?

- Is there any missing or incomplete data?

This information can be obtained by using info() command. Although, further analysis could be required.

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

Yes, except the dropoff and pickup datetime needs to be checked if it's in proper format

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

First of all, know your data, the size, the type of data and the info included in the dataset

PACE: Analyze Stage

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

Once you are familiar with the dataset, you can later identify if there are any outliers with a box plot or histogram.

Once you identified it, you make the decision of what to do with these data before doing any analysis

Finally, you perform EDA and find any useful insight between variables or how they relate them

- 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.?

Since it's only one dataset, joining or merging is not required for this process.

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

The visualization needs to be adequate for any audience, and the graphics needs to be clear and concise.

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 can create visualizations about the revenue by month or day. Also, It could be important to create graphics about the ride count per day or month

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

It depends on the data. Sometimes you need to make sure the data its in proper format, then filter it if its necessary and perform the proper calculations or actions.

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

Pickup and dropoff datetime, total amount and trip distance

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

First of all, using info() command to see the non-null values and then perform box plot, histogram and scatter plot.

PACE: Execute Stage

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

Total revenue and Ride count per day and month. Also, a scatter plot for the Total Amount and Trip Distance

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

In some data, like the trip duration, there is data for 0 which it doesn't makes sense. Also, there was some outliers in the data but it didn't interfere with the EDA

- 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 ask them if they have the coordinates for the

- How might you share these visualizations with different audiences?

Putting the right colors for any audience, and making the visuals clear and easy to read.