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

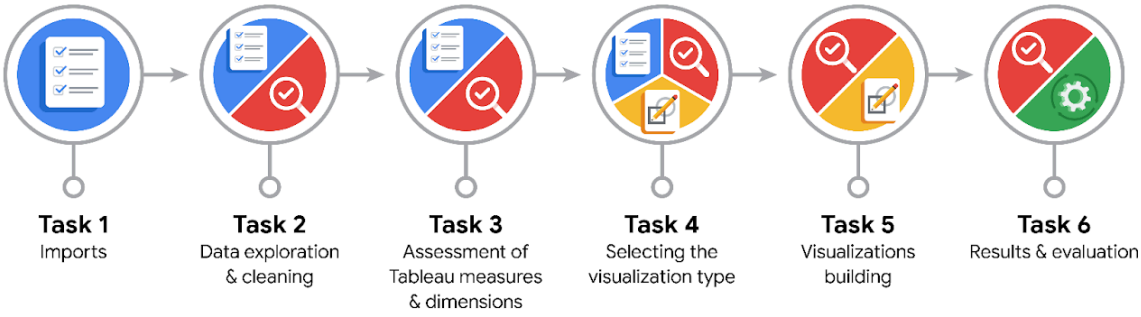
Relevant Interview Questions

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?

Trip_distance, total_amount are the two major columns that will impact the predictions on the estimated fare.

- What units are your variables in?

float64, int64 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?

That there are outliers that might impact the model, if not dealt with.

- Is there any missing or incomplete data?

No

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

Yes

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

We will begin by discovering and structuring the data. We will also present the data. Then carry on with the rest of the EDA practices while presenting the data in between where we feel like it might be helpful to the stakeholders.

PACE: Analyze Stage

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

Discovering, structuring, cleaning, and presenting

- 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, we do not need to add more data for our ultimate goal. Sorting, filtering, extracting, grouping are the major structuring that needs to be done to this dataset.

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

Boxplot, histograms, bar graphs and scatter plot

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?

Boxplots and histograms (for detecting outliers), bar graphs (for comparison of the relationship between two variables), scatterplot (for identifying trends).

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

We need to extract the data, clean it and make it ready to graph it.

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

Trip_distance, total_amount are the two major variables that are most applicable for the visualizations in this data project.

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

There are 3 ways - i) Ignore them, ii) Fill the average value of the non-missing values in place of the missing value and, iii) Observe the dataset to see if there are any relation or trends between the missing and non-missing values.

PACE: Execute Stage

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

The highest distribution of trip distances are below 5 miles, but there are outliers all the way out to 35 miles. There are no missing values. The majority of trips were journeys of less than two miles. Nearly two thirds of the rides were single occupancy, though there were still nearly 700 rides with as many as six passengers. Also, there are 33 rides with an occupancy count of zero, which doesn't make sense.

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

The data includes dropoff and pickup times. We can use that information to derive a trip duration for each line of data. This would likely be something that will help the client with their model.

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

Ride count by hour of the day. I can also join the dataset with a holiday list and see how the ride count fluctuates during holidays.

- How might you share these visualizations with different audiences?

Use suitable colors for the color blind people, contrast, appropriate titles and labels, add Alt text, and make interactive dashboards for the non-technical audience to get the visualizations easily.