# 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

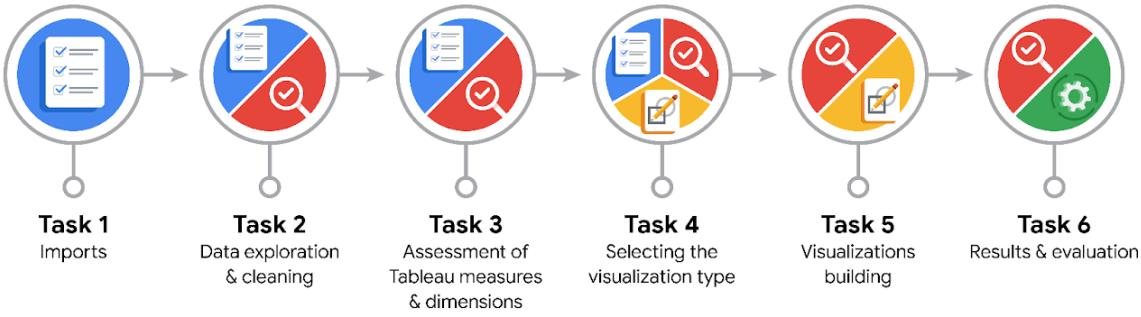
### Relevant Interview Questions

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?

| |
|---|
| In this dataset, trip distance and total amount appear to be useful columns to analyze along with combining others. |

- What units are your variables in?

| |
|---|
| Integer, object, and float are the units in all variables of the dataset. |

- 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 in data for example, which in the case of collecting data. |

- Is there any missing or incomplete data?

There is no missing data in the dataset.

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

For all variables, there are included strings, numerical, and categorical.

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

Using the six practices of EDA: discovering, structuring, cleaning, joining, validating, and presenting.

## PACE: Analyze Stage

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

The first is discovering, which is to help understand the general of the whole dataset. Structuring data is to deal with every single variable to further examine and manipulate them. Cleaning all the missing values, duplicates, inconsistent format, etc. in variables. We might consider joining if there's a need to combine with another dataset. Validating is to recheck/double-check what we've done before to ensure the data is high-quality, and error-free. Presenting is to deliver the key results, insights, and findings to stakeholders, and recommend the next steps to take.

- 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. Many types of structuring can be used such as filtering, sorting,

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

Box plots, scatterplots, bar plots, and histograms might be best suited for 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?

In this project scope, visualizations are the most essential. Box plots, histograms, bar graphs, and scatterplots are visualized in this dataset to examine important variables.

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

First by determining the variables' type. In the case of the univariate analysis, histograms and box plots are the best approach. Bivariate analysis might consider bar graphs or scatter plots with relationships between variables. Multivariate analysis is to examine more variables, such as using a pair plot.

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

The most two important variables are trip distance and total amount, which are used consistently to visualize.

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

- There are many ways to decide if there are missing values:
    - Choose to drop out in case the percentage isn't too much or minimal.
    - If there's missing too much, consider contacting the owner to request filling in the missings.

- Choose to impute the missings. For categorical, we can group the missings as **answers not recorded**, or something else relevant. For numerical, impute the mean or median.

## PACE: Execute Stage

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

Many variables such as trip distance, total amount, and tip amount appear as right-skew distributions. The tip amount for every passenger count is pretty low <$2. The months of March-June appear to have consistently on several rides and the total revenue. For weekdays, most rides and total revenue are on the mid-week. The relationship between trip distance and total amount is positive. However, there are trips with 0 distance at all but still have the total amount.

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

First, we might address the issue of trip distance and total amount. Then, from the insights of each visualization, we might recommend the decisions to take or want to further analyze. Plus, from the client's perspective, they might want to know whether any other variables might be effective in building the predictive model of ride amount.

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

With the scatter plot compared between trip distance and total amount, how likely is it possible to have 0 distance but still have the amount? Is there a problem while collecting data, errors, or something else related?

- How might you share these visualizations with different audiences?

| |
|---|
| Within the data team, it's the best practice to go over the bunch of visualizations and explain key insights, and deeper details. With non-technical audiences, a simple visualization, detailed explanation, and account for visual impairment will be effective. |