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

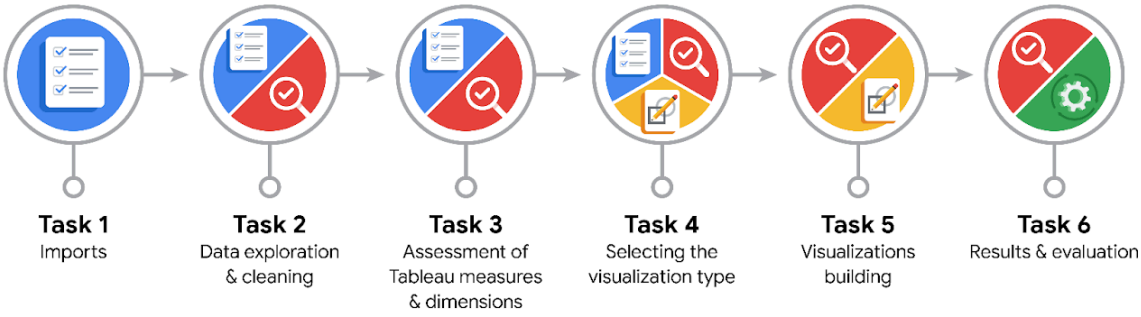
Relevant Interview Questions

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?

Data columns and variables relevant to my deliverable would include trip distances, drop off and pick up times and locations, total costs, and tip costs.

- What units are your variables in?

Times and location variables are in datetimes. Revenue and costs are in USD.

- 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?

My initial presumptions can be confirmed because the fare costs are positively correlated with the trip distances, and also there are outliers I got to deal with here before modelling stuff.

- Is there any missing or incomplete data?

There are no missing data.

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

No, I have to convert the date times to a datetime variable.

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

Dealing with outliers, visualizing the data, and make sure every types are correct.

PACE: Analyze Stage

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

Data cleaning and visualization needs to be taken to perform EDA.

- 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.?

Yeah, filtering and sorting could be a way to add more data using the EDA practice of joining.

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

Bar plots for comparing trip averages by month or weekday, scatterplots for overall comparison, and boxplots/histograms for visualizing a single variable, such as trip distances and total fares.

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 section, creating histograms, boxplots, bar plots, and scatterplots using Python or Tableau would be relevant in this goal.

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

Using the appropriate libraries such as matplotlib and seaborn and then filtering out the appropriate data would be needed before building the visualization.

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

Total cost, tip cost, pick up times, drop off times, locations, (for datetime variables specifically the month of year and day of week) are the most applicable for visualizations in this step.

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

PACE: Execute Stage

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

Key insights include there are outliers. Appropriate methods would need to be dealt.

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

I would recommend, based on the visualizations, that outliers to be treated appropriately and to be re-computed or kept when doing hypothesis testing and modelling.

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

I can research: does the ride distances and fares depend on the day of the week? The same can also be asked depending on the month of the year.

- How might you share these visualizations with different audiences?

For the people that does not understand data analytics too much, I will try to use words as much as possible to describe the visualizations.