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

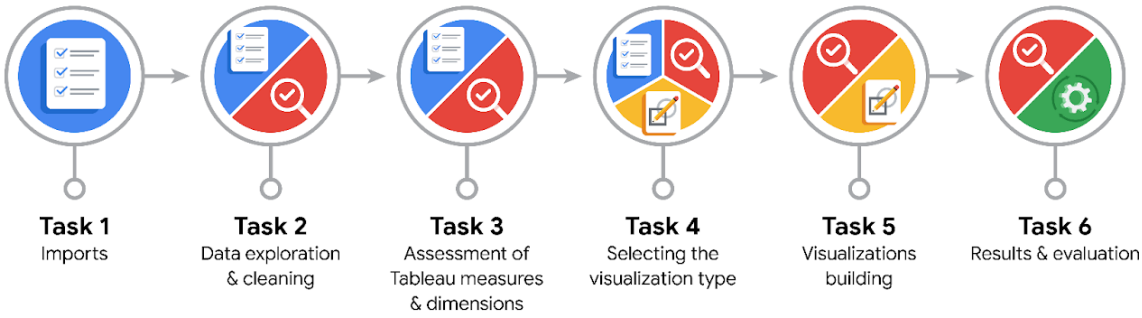
Relevant Interview Questions

Completing the end-of-course project will help you respond 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?

The dataset consists of 18 columns including variables like trip_distance, total_amount etc.

Most relevant columns : trip_distance, total_amount, **pickup_datetime** and **dropoff_datetime**:

- What units are your variables in?

Miles and Dollars

- 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?

Strong correlation between the number of rides and revenue.

- Is there any missing or incomplete data?

There is no missing Data

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

8 columns have dtype float, 7 have int and 3 have object

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

Importing libraries and loading the 2017 New York TLC Dataset

PACE: Analyze Stage

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

Clean the data, remove outliers, and explore key metrics like trip_distance, and total_amount using summary statistics and visualizations.

- 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.?

Not necessary for this dataset. Focus on filtering and sorting relevant columns for analysis.

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

Boxplots and histograms

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?

Box plots, Histograms and Bar charts. Also, scatterplots in tableau to help showcase presence of outliers

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

Data Exploration and cleaning

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

trip_distance, passenger_count, total_amount etc

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

No missing data. Missing data was dealt in prior labs

PACE: Execute Stage

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

Short Trips Dominate. Single Passenger rides are more than 2/3. Peak times can be pinpointed.

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

Deeper analysis on shorter trips, optimize services in high-demand areas, conduct hourly rides breakdown by frequency and revenue to calculate peak times and adjust prices accordingly.

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

Investigate efficiency and profitability of different routes

- How might you share these visualizations with different audiences?

Present interactive dashboards for business users, and reports with detailed visuals for stakeholders.