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

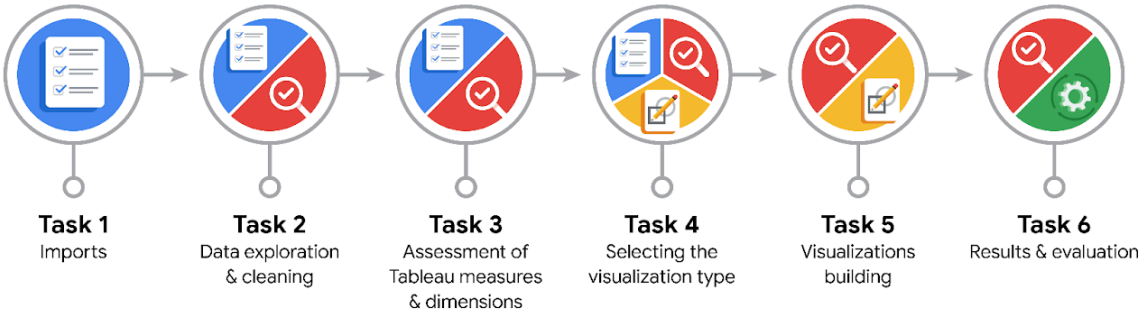
Relevant Interview Questions

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 duration and total amount are the most relevant variables to the deliverable.

- What units are your variables in?

Trip duration is in miles and total amount is in dollars (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?

I assume that longer trips correspond to longer fares; data that doesn't reflect this trend must be addressed.

- Is there any missing or incomplete data?

  The number of rows for the dataset does not correspond to the number provided by TLC. The current rows do not have any null values.

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

  The columns have different datatypes: numeric data is stored in int64 and float64, and there are two column of object data type.

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

  Data cleaning, data wrangling, and visualizations are required to understand relationships between the variables and to find any discrepancies in the data.

PACE: Analyze Stage

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

  Any outliers that might negatively impact model performance must be removed. Features that have a significant correlation or importance to the target must be discovered while insignificant features are to be ignored.

- 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.?

  Additional features can be created through encoding categorical variables. Low cardinality features can be one-hot encoded and high cardinality features can be target encoded.

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

Boxplots can be used to show outliers and histograms can show distributions of data that display trends relevant to the target.

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, bar charts, scatterplots, and histograms are needed to visualize data patterns. Normalization and random-forest regression can be machine learning algorithms that guide model training.

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

Visualizations can be built using Python via the seaborn and matplotlib libraries, or through Tableau Public dashboard software.

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

Trip duration, total amount, DOVendorID, DOdate are applicable for visualizations.

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

Missing data can be imputed with the average value if continuous or dropped if the missing data count is low.

PACE: Execute Stage

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

There are trip durations of 0 for positive total amounts. Trip duration and total amount appear to be the most significant features in relation to the target.

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

Remove outliers from the dataset and discard insignificant features for model training.

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

How does location affect fare amount? What factors lead to higher tips for vendors?

- How might you share these visualizations with different audiences?

Non-technical stakeholders can interact with Tableau dashboards to have a better understanding of variable relationships. Technical stakeholders can look at heatmaps and distribution plots.