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

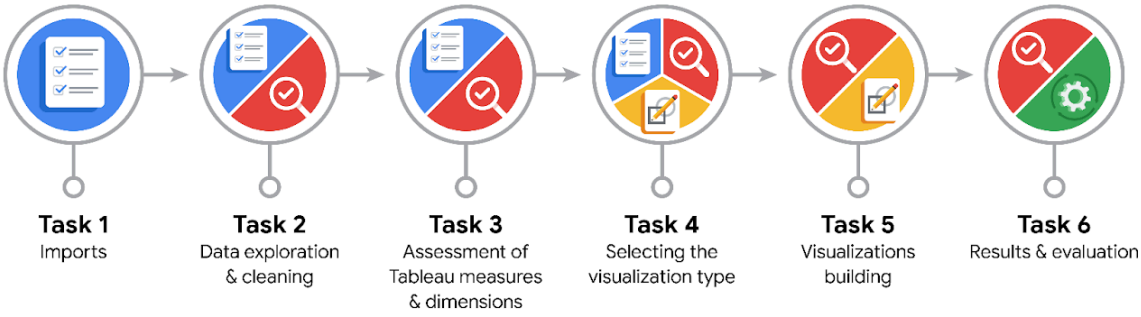
Relevant Interview Questions

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?

The dataset includes columns like VendorID, tpep_pickup_datetime, tpep_dropoff_datetime, Passenger_count, Trip_distance, PULocationID, DOLocationID, RateCodeID, Payment_type, Fare_amount, Tip_amount, Tolls_amount, and Total_amount. For estimating taxi fares, Trip_distance, PULocationID, DOLocationID, RateCodeID, Passenger_count, Payment_type, and Total_amount are most relevant.

- What units are your variables in?

Trip_distance is in miles. Fare_amount, Tip_amount, Tolls_amount, and Total_amount are in USD. tpep_pickup_datetime and tpep_dropoff_datetime are in datetime format.

- 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 expect Trip_distance to have a strong positive correlation with Total_amount. I also expect RateCodeID to influence Total_amount significantly. I anticipate that trips during peak hours may have higher fares.

- Is there any missing or incomplete data?

An initial check is required to identify missing values in columns like Passenger_count or any of the dollar amount columns. Also the location ID columns should be investigated.

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

The datetime columns need to be verified for consistency. PULocationID and DOLocationID should be checked to ensure they are consistent with the TLC zone definitions. The Payment_type column should be checked to make sure that the numeric codes match the data dictionary.

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

Descriptive statistics, histograms, scatter plots, box plots, and correlation analysis will be crucial. We also need to perform data type checks and handle missing values. So, the required EDA practices to begin with, are Discovering, Structuring, Cleaning, and Validating (as an iterative process).

PACE: Analyze Stage

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

First, I'll examine the distribution of Trip_distance, Total_amount, and Passenger_count. Then, I'll analyze the relationships between these variables and RateCodeID and Payment_type. Finally, I'll explore the spatial distribution of trips using PULocationID and DOLocationID.

- 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.?

  Joining with a TLC zone lookup table could provide more meaningful location information. Filtering by RateCodeID and sorting by Total_amount will be necessary. Creating new columns for trip duration and time of day could be useful.

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

Scatter plots for Trip_distance vs. Total_amount, bar charts for Payment_type distribution, and maps showing trip origins and destinations would be effective. Box plots to show the distribution of fares for each RateCodeID would also be useful.

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?

Scatter plots, bar charts, box plots, histograms, and maps. We will be constructing a regression model, but for this part of the project we will focus on the visualizations.

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

Data cleaning, feature engineering (e.g., trip duration), grouping, aggregation, and then using libraries like Matplotlib, Seaborn, and Tableau to create the charts and maps.

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

Trip_distance, Total_amount, RateCodeID, Payment_type, PULocationID, and DOLocationID are the most applicable.

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

Missing Passenger_count values might be imputed with the mode. Missing dollar amount values will need to be investigated, and rows with those missing values may need to be removed. Missing location ID's will need to be investigated, and any rows with those missing values may also need to be removed.

PACE: Execute Stage

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

This dataset has some outliers that we will need to make decisions on prior to designing a model. 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 drop-off points are relatively evenly distributed over the terrain. Strong positive correlation between Trip_distance and Total_amount. Certain pickup and dropoff locations have higher trip volumes. Trip duration and time of day appear to be strong factors in total fare amount

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

The TLC can use the regression model to provide fare estimates to passengers. They can also analyze high-volume locations to optimize taxi availability. Further analysis of the rate codes can be performed to see if the rates are still optimal. They can use the payment type data to help make determinations about how to improve payment processing.

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

What are the specific factors contributing to high trip volumes in certain locations? How does traffic congestion affect trip duration and fare? Are there seasonal trends in taxi ridership? How can the regression model be improved with additional features?

- How might you share these visualizations with different audiences?

For the Automatidata team, provide detailed reports with code and statistical analysis. For the TLC team, create an interactive dashboard with clear and concise visualizations, focusing on key insights and recommendations. Use non-technical language and focus on the business implications.