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

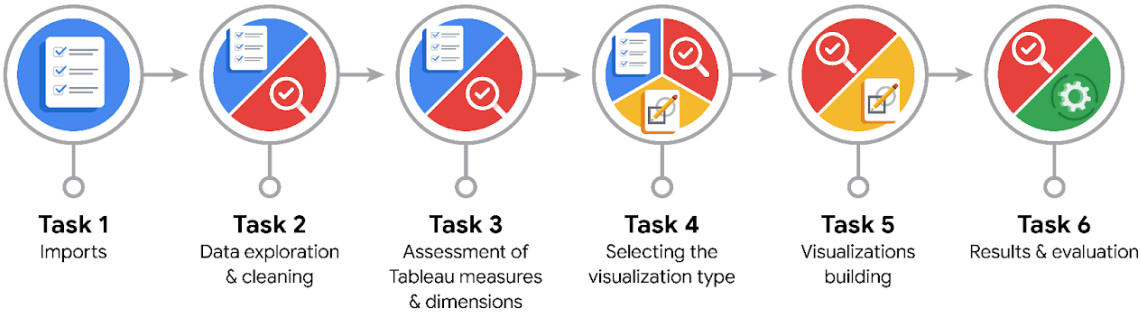
Relevant Interview Questions

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 data columns include VendorID, pickup_datetime, dropoff_datetime, passenger_count, trip_distance, RatecodeID, store_and_fwd_flag, PULocationID, DOLocationID, payment_type, fare_amount, extra, mta_tax, tip_amount, tolls_amount, improvement_surcharge, total_amount. Relevant columns: pickup_datetime, dropoff_datetime, passenger_count, trip_distance, PULocationID, DOLocationID, payment_type, fare_amount, tip_amount, total_amount.

- What units are your variables in?

Most variables are in common units: trip_distance: miles fare_amount, tip_amount, total_amount: US dollars pickup_datetime, dropoff_datetime: datetime format passenger_count: count of passengers

- 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?

There might be seasonal or time-based trends in ride counts and tips. Passenger count might influence the tip amount. Certain locations might have higher or lower trip distances and fares.

- Is there any missing or incomplete data?

Check for any missing values in key columns like pickup_datetime, dropoff_datetime, trip_distance, total_amount, etc.

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

Ensure datetime columns are in datetime format. Ensure numerical columns are in appropriate numeric formats.

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

Data cleaning (handling missing values, correcting data types) Descriptive statistics Data visualization (histograms, box plots, bar charts, scatter plots)

PACE: Analyze Stage

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

Load and clean the data Convert columns to appropriate data types Identify and handle missing values Perform descriptive statistics

Create initial visualizations to understand data distribution

- 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.?

No additional data joins required initially Structuring tasks: sorting by date, filtering out invalid values, creating new columns (e.g., month, day)

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

Histograms for distribution analysis Box plots for identifying outliers Bar charts for categorical comparisons Line charts for time series analysis Scatter plots for relationship analysis

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?

Data visualizations: Histograms, box plots, bar charts, line charts, scatter plots Potential ML models: Regression analysis to predict trip distances or fares based on input features

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

Data cleaning and preprocessing Calculating descriptive statistics Creating new features (e.g., month, day of week) Plotting the visualizations using libraries like Matplotlib and Seaborn

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

trip_distance, tip_amount, total_amount, passenger_count, pickup_datetime, dropoff_datetime, PULocationID, DOLocationID

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

Handle missing data by either filling with appropriate values (e.g., mean/median) or dropping the rows/columns if the missing data is significant.

PACE: Execute Stage

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

Seasonal and time-based trends in ride counts and tips Higher tips associated with certain passenger counts Significant differences in trip distances and fares based on pickup and drop-off locations

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

Focus marketing and promotions on peak times identified through EDA Encourage tipping by highlighting patterns in passenger counts and service quality Optimize routes based on high-demand pickup and drop-off locations

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

Investigate the impact of external factors like weather or special events on ride demand Analyze driver performance and efficiency Explore customer satisfaction metrics and their correlation with tips and ride frequency

- How might you share these visualizations with different audiences?

Present high-level summaries and key insights for executives
Provide detailed visualizations and technical explanations for data analysts and technical teams
Use clear, accessible charts and narratives for broader organizational stakeholders