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

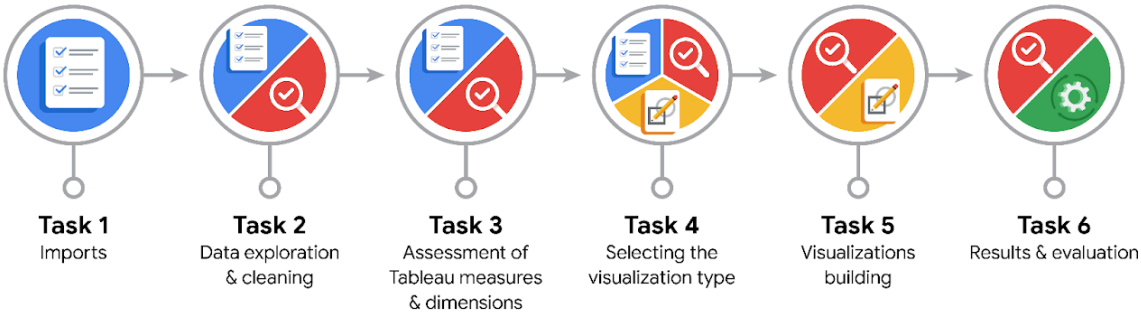
Relevant Interview Questions

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?

Relevant columns: `tpep_pickup_datetime`, `tpep_dropoff_datetime`, `passenger_count`, `trip_distance`, `PULocationID`, `DOLocationID`, `fare_amount`, `tip_amount`, `total_amount`. These are relevant because they provide information related to trip characteristics, revenue, and location.

- What units are your variables in?

`trip_distance`: miles `fare_amount`, `tip_amount`, `total_amount`: US dollars `tpep_pickup_datetime`, `tpep_dropoff_datetime`: date and time format `passenger_count`, `PULocationID`, `DOLocationID`, `VendorID`, `RatecodeID`: numerical IDs.

- 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?

• Trip distances and fare amounts will be right-skewed. • Ridership will vary by day of the week and month. • Certain pickup and drop-off locations will have higher traffic. • Tip amount will correlate with fare amount.

- Is there any missing or incomplete data?

Based on the `info()` function, there are no missing values in this dataset.

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

Most numerical data is in `int64` or `float64` format. Date/time columns are `object` and need to be converted to datetime.

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

1. Data cleaning (datetime conversion). 2. Descriptive statistics (`describe()`). 3. Visualization (box plots, histograms, bar charts, scatter plots). 4. Grouping data (groupby()).

PACE: Analyze Stage

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

1. Convert date/time columns. 2. Identify and handle outliers. 3. Analyze distributions of key variables. 4. Explore relationships between variables. 5. Visualize trends over time (month, day).

- 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.?

For this initial EDA, no additional data is needed. Structuring: - Filtering outliers. - Grouping by date/time components, location IDs, and passenger count. - Sorting by date/time, distance, and revenue.

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

- Bar charts for comparing categorical data (e.g., rides per month/day). - Box plots and histograms for distribution analysis. - Scatter plots (in Tableau) to show relationships between continuous variables. - Geographic map (in Tableau) to show location data.

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, bar charts, and scatter plots. Geographic map in Tableau.

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

Data cleaning and transformation. Using `matplotlib` and `seaborn` for Python visualizations. Using Tableau Public for interactive visualizations.

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

`trip_distance`, `fare_amount`, `tip_amount`, `total_amount`, `tpep_pickup_datetime`, `DOLocationID`, `passenger_count`

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

As stated before, there is no missing data in this dataset. If there were missing data, the plan would be to remove rows if the quantity was low, or impute the data if there was a large amount of missing data.

PACE: Execute Stage

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

Outliers are present in trip distance, fare, and tip amounts.
Ridership and revenue vary by month and day of the week.
Certain drop-off locations have higher average trip distances.
Tip amounts are higher with higher passenger counts.

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

Investigate outliers for potential fraud or data errors.
Optimize staffing and resources based on peak hours and days.
Identify high-demand areas for targeted marketing or resource allocation.

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

How do weather conditions or major events affect ridership?
What are the reasons for the observed outliers?
Can we predict future ridership or revenue based on historical data?
What are the most common trip routes?

- How might you share these visualizations with different audiences?

Executive summary with key findings and recommendations.
Interactive Tableau dashboard for detailed exploration.
Python notebook for technical audiences.
When presenting to visually impaired people, be sure to use high contrast colors, and very descriptive titles, and axis labels. When possible, provide the underlying data in a table format.