**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 includes variables like pickup\_datetime, dropoff\_datetime, trip\_distance, passenger\_count, fare\_amount, tip\_amount, and vendor\_id. Key variables for this project include ride duration, trip distance, pickup/drop-off locations, and date/time features.

* What units are your variables in?

Time variables are in timestamps (datetime format)

Distance is in miles

Fare and tip amounts are in U.S. 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?

I initially presumed that longer trips would lead to higher fare amounts and that weekends or certain months might show increased ridership. I also assumed that trip patterns could reveal peak hours and geographic demand differences.

* Is there any missing or incomplete data?

Yes. Some records had missing or null values for trip distance, fare, or timestamps. These were filtered or imputed as needed.

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

No. Some variables (like dates) needed conversion from string to datetime, and some monetary values needed to be cleaned or formatted consistently.

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

Basic EDA practices included:

* Checking data types
* Detecting and handling missing values
* Identifying outliers using box plots
* Time-based aggregations
* Univariate and bivariate analysis

**PACE: Analyze Stage**

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

Clean and standardize the data

Convert datatypes appropriately

Create new columns such as ride duration

Aggregate rides by month and vendor

Visualize trends over time and distributions of key variables

* 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 datasets were added. Structuring included filtering out invalid or zero-distance trips, sorting by date/time, and creating datetime features (month, quarter).

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

For management and stakeholders:

* **Box plots** to show ride duration distributions
* **Line charts** or **time series plots** for monthly trends
* **Bar charts** to compare vendors or tips by distance
* **Map visualizations in Tableau** for geographic insights  
  Accessibility considerations were made for color contrast and simplicity.

**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?

No machine learning was used. Visualizations included:

* Box plot of ride durations
* Time series plots (monthly, quarterly trends)
* What processes need to be performed in order to build the necessary data visualizations?

Clean data and remove anomalies

Group data by desired categories (month, vendor, etc.)

Aggregate metrics like average fare or tip per trip

Plot charts using Matplotlib/Seaborn

* Which variables are most applicable for the visualizations in this data project?
* pickup\_datetime, dropoff\_datetime
* trip\_distance
* fare\_amount, tip\_amount
* passenger\_count
* PULocationID, DOLocationID
* Going back to the Plan stage, how do you plan to deal with the missing data (if any)?

Rows with critical missing values (e.g., fare amount, timestamps) were dropped. For non-critical missing fields, default or mean-based imputation was considered.

******PACE: Execute Stage**

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

Rides peak during summer and early fall.

Longer trip durations are associated with higher fares but show more variability in tips.

Most trips are short in distance and clustered within city boroughs.

Vendor differences in average tip amounts may indicate customer preference or quality differences.

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

Focus marketing efforts during peak ride months (e.g., July–October).

Explore reasons behind vendor performance differences to improve consistency.

Optimize fleet deployment in high-demand zones at peak hours.

Standardize tipping prompts or options to improve revenue predictability.

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

How does weather influence ride duration or frequency?

What role do holidays play in ridership spikes?

Is there a correlation between specific pickup zones and tipping behavior?

* How might you share these visualizations with different audiences?

**For executives**: Present summarized Tableau dashboards highlighting trends and KPIs.

**For analysts**: Share Python notebooks with full EDA and code explanations.

**For public stakeholders**: Use clean, accessible visual dashboards with descriptive labels and tooltips, adhering to visual accessibility standards.