**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?

There are numerous data columns associated with a single ride. However, the most important variables, I believe, happen to be Trip Distance, Total Amount and Tip Amount.

* What units are your variables in?

**Trip Distance**: Measured in miles

**Trip Duration**: Measured in minutes or seconds.

**Fare Amount**: Measured 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?

Longer trips are likely to result in higher fares, assuming other factors remain constant.  
Fares may vary depending on whether the trip occurs during peak hours, at night, or on weekends, where additional charges (e.g., night surcharges) could apply.

* Is there any missing or incomplete data?

The dataset is complete, and has no incomplete data.

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

Date columns had to be changed to DateTime format.

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

Understand the Structure of the Data, Handle Missing or Incomplete Data, Outlier Detection, Examine Distributions.

**PACE: Analyze Stage**

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

Familiarize yourself with the dataset's overall structure and content.  
Identify and address missing or incomplete data to ensure the dataset is usable. Identify and assess the impact of outliers that could skew the model.

* 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.?

Add historical weather data for the time of each trip, as weather conditions can affect demand and fares. Include data on local events or holidays that might influence taxi demand and fares (e.g., concerts, sports events). Access historical traffic conditions that could impact travel times and fare calculations.

#### **Steps to Join Data:** Use pd.merge() to combine your taxi dataset with the additional datasets on common keys (e.g., date, location).

Ensure that the additional data aligns correctly with the existing dataset and that data types are consistent.

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

**Histograms**: To show the distribution of continuous variables like fare amounts, trip distances, and trip durations. This will help identify skewness, outliers, and the general shape of the data distribution.

**Box Plots**: To visualize the distribution of fares across different categories, such as vendor\_id or payment\_type. This can highlight outliers and the interquartile range.

**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?

**Exploratory Data Analysis (EDA) Visualizations**

Histograms: Visualize the distribution of continuous variables (e.g., fare amounts, trip distances).

Box Plots: Compare fare distributions across categories (e.g., vendors, payment types) to identify outliers.

Heatmaps: Show correlations between numerical variables to aid feature selection.

Time Series Line Charts: Examine trends in fares by time (hour, day of the week) to identify peak usage.

**Regression Algorithms**

**Model Evaluation Techniques**

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

**Data Preparation**: Data Cleaning, Outlier Detection, Data Type Conversion, Subsetting the Data, Aggregation

**Selecting Visualization Types**

**Creating Visualizations**

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

Trip Distance and Total Amount

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

Use methods to check for missing values in the dataset. ssess the percentage of missing values for each variable. If a variable has a high percentage of missing data, it may not be useful for modeling. Analyze whether the missingness is random (Missing Completely at Random, MCAR), not at random (MNAR), or missing at random (MAR), as this will influence how you address it.

Decide on a Strategy for Missing Data: Remove or keep it

Evaluate the Impact of the Chosen Method

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

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

There are instances where the trip distance is 0 but they have fare amounts.

There are negative fare amounts.

There are outliers with fare amounts as high as $1200.

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

Determine any unusual data points that could pose a problem for future analysis in predicting trip fares. For example, locations that have longer durations.

Determine the variables that have the largest impact on trip fares.

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

Trips that have a total cost entered, but a total distance of ‘0’ need to be questioned about. Were those rides canceled and waiting costs incurred?

* How might you share these visualizations with different audiences?

#### **For Technical Stakeholders**

**Interactive Dashboards**: Use tools like Tableau, Power BI, or Plotly Dash to create interactive visualizations that allow users to explore the data on their own.

**Jupyter Notebooks**: Share Jupyter Notebooks that include code, visualizations, and explanations. This allows for reproducibility and deeper technical engagement.

**Documentation and Reports**: Provide comprehensive reports with detailed methodology, data sources, and technical explanations of visualizations and findings.

#### **For Business Stakeholders**

**PowerPoint Presentations**: Create concise presentations summarizing key findings with clear visualizations. Use bullet points and straightforward language to convey insights effectively.

**Summary Reports**: Provide high-level reports with an executive summary, focusing on actionable insights, implications for the business, and recommendations.