**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 three most relevant columns are the trip pickup, trip dropoff, and total amount columns.

* What units are your variables in?

The variables are in datetime (trip pickup and dropoff) and total amount (int64) 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 presume we will find that on average the greater the distance, the higher the cost of a ride.

* Is there any missing or incomplete data?

There is no missing data.

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

No. There are integer, string, and datetime formats.

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

Converting the Pickup and Dropoff columns into datetime objects.

**PACE: Analyze Stage**

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

Arranging the columns in order.

Creating new columns for month and day from the pickup and dropoff columns.

Creating box plots and bar charts to identify outliers.

* 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 more data will be needed.

The data needs to be filtered to focus on the most important information, and sorted according to time and greatest price amount.

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

Scatter Plots should be the best type of visualization, as they will show where the majority of data points connect between the two variables. Scatter Plots will also be useful to show outliers from the pattern.

**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 will be needed to complete the EDA to provide stakeholders with a good visualization of the data.

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

The dataset needs to be validated, uploaded to Tableau, and filtered to highlight the relevant variables.

* 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)?

By filtering out null values.

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

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

The key insights gained are that certain fare amounts registered at less than zero, while one fare amount was far higher than the rest. This suggests an outlier that needs identified.

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

Focus the regression model on the correlation between trip distance and fare amount.

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

Why are there negative outliers?

Is there a correlation between trip time and fare amount?

* How might you share these visualizations with different audiences?

Tableau Public would be a great way to share these visualizations, as well as Kaggle.