**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 most relevant columns to our deliverables are: Vendor ID, payment type, trip distance, fare amount, tip\_amount, total\_amount. For a predictive model, the most useful columns are trip distance and total amount. Trip duration (which needs to be created) may be relevant as well.

* What units are your variables in?

Trip distance is in miles. Other relevant columns are in dollars. Vendor ID and payment type are identifiers without a specific unit.

* 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 are several irrelevant columns. We have outliers that need to be evaluated for potential cleaned-up.

* Is there any missing or incomplete data?

Missing data appears to be limited. There are zero entries in fare amount that likely we need to deal with.

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

The relevant columns in the dataset are in the same format

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

Discovering, Validating, Structuring, Validating, Cleaning-up, Presenting.

**PACE: Analyze Stage**

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

Structuring. Evaluate up for outliners and missing data. Validating the data is in same format and usable. Then presenting.

* 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.?

Joining is not required. Filtering, grouping and sorting help in visualizations. Creating a new column for trip duration as part of structuring.

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

Box plot will highlight outliers. Histogram and bar charts will be useful to present the 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 plot, histogram and bar charts. In next steps, we would need a predictive model.

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

Structuring, Cleaning-up.

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

Trip distance and total amount (or fare amounts). Trip duration may be useful.

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

Most likely, missing data should be deleted from the data set. This will need to be confirmed.

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

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

The majority of trips are in the 1-3 miles range.

Most total amounts paid for fares fall in the $5-$20 range. Tips are in the $0-$3 range.

Wednesday through Saturday had the highest number of daily rides, while Sunday and Monday had the least.

Thursday had the highest gross revenue of all days, and Sunday and Monday had the least. Interestingly, although Saturday had only 35 fewer rides than Thursday, its gross revenue was ~$6,000 less than Thursday's—more than a 10% drop.

Monthly revenue generally follows the pattern of monthly rides, with noticeable dips in the summer months of July, August, and September, and also one in February.

Notice that out of the 200+ drop-off locations, a disproportionate number of locations receive the majority of the traffic, while all the rest get relatively few trips. It's likely that these high-traffic locations are near popular tourist attractions like the Empire State Building or Times Square, airports, and train and bus terminals.

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

It would be helpful to know the location that each ID corresponds with. Unfortunately, this is not in the data.

The highest distribution of trip distances are below 5 miles, but there are outliers all the way out to 35 miles. There are no missing values.

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

There are several trips that have a trip distance of "0.0." What might those trips be? Will they impact our model?

The data includes dropoff and pickup times. We can use that information to derive a trip duration for each line of data. This would likely be something that will help the client with their model.

* How might you share these visualizations with different audiences?

Consider the more and less technical audiences to decide what/how to share with them.