**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?

For me, it is trip\_distance, tip\_amount, total\_amount, passenger\_count columns

* What units are your variables in?

Integer and Float

* 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?

For me this is a complex dataset and need exploring carefully.

* Is there any missing or incomplete data?

There is no missing data

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

No. There’re int, float and object datatypes.

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

I will discover the dataset first by having a look at the types of data, basic statistical summaries, and their distribution visualizations.

**PACE: Analyze Stage**

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

For me, it is cleaning and preprocessing the dataset. It’s like cooking, it takes you a lot of time to prepare the necessary ingredients. Having a structured dataset is ideal to achieve project goal.

* 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.?

Yes. I need PULocationName, DOLocationName to understand the PULocationID and DOLocationID variable. We can use sorting by day, week, month and LocationID to understand more the characteristics of the dataset.

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

Bar plot and histogram.

**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?

Data visualizations: Bar charts, line charts, pie charts, heatmaps, and other visualizations can be used to show the distribution of data, trends over time, and the relative proportions of data.

Machine learning algorithms: Machine learning algorithms such as linear regression can be used to predict future trends, identify patterns in data, and make recommendations.

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

Filtering and grouping the data

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

Variable tip\_amount and total\_amount

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

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

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

The more trips in day/week/month, more revenue it is.

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

Increase the number of rides during the off-peak hours. The visualizations show that the busiest times of day for taxi rides are in the morning and evening rush hours. By increasing the number of rides during the off-peak hours, the taxi company can increase its profits.

Optimize the taxi routes. The visualizations show that the taxi company's routes are not always optimized. By optimizing the routes, the taxi company can reduce the distance traveled by its taxis and save money.

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

What’s the average trip duration?

* How might you share these visualizations with different audiences?

I would use Tableau to perform a whole story