**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 relevant columns are: trip\_distance, fare\_amount. Relevant variables total revenue, which can be easily calculated from available data.

* What units are your variables in?

Trip\_distance is miles and fare\_amount is 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?

There is probably a relationship between income and days of week and months. Additionally,

* Is there any missing or incomplete data?

Inicial data cleaning shows that there is no missing data, however there are some strange looking outliers.

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

No, there are: int, float and object.

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

Identifying trends between data, finding outliers and potential errors.

**PACE: Analyze Stage**

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

Discover if amount of data is enough for effective study, validate if data has no errors, clean the data and validate if cleaning was successful.

* 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.?

I think provided data will be sufficient for research. The things that need to be done are: data cleaning, filtering only relevant data, sorting it by day/month.

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

The best way would be to use several types of plots to visualize the data. For example, distance would be best shown on box or scatter plots while revenue depending on month or day on bar charts.

**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?

Ways to easily visualize data such as: revenue and trip distance. Also, it would be best to find a way to easily discard false data.

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

First cleaning the data, then filtering the data, finally sorting.

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

Trip distance and revenue by day/month.

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

First thing is if there is a connection between missing data (like everything comes from one source/place) I would investigate to check if data was in some way corrupted and can be retrieved. Otherwise, if deleting missing data would not have massive impact on outcome of the research I would just do it.

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

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

Vast majority of people take a trip of less than 2 miles, the cost of most of them falls between 5 and 15$. Number of rides is mostly consistent with the exception of dips seen in February, July and August. Monday and Sunday have the least daily rides while Wednesday has the most.

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

To reduce upkeep costs it would be advised to have a reduced number of ready taxis (or reduce working hours) on Mondays and Sundays (except special holidays) while increasing the number on Wednesday. Same goes for months - having less taxis in months with less rides would reduce costs.

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

Which vendor is responsible for most revenue, what is the most chosen payment method (and does it influence the amount paid, tip etc.), what are the peak hours for taxis.

* How might you share these visualizations with different audiences?

By using Tableau - while plots generated by python would most probably be readable by data professionals, it makes data visualization much easier for a casual audience.