**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. As a reminder, this document is a resource that you can reference in the future and a guide to help consider responses and reflections posed at various points throughout projects.

# Course Project Recap

Regardless of which track you have chosen to complete, your goals for this project are:

* Understand and assess the proposed scenario
* Demonstrate understanding of how to organize and analyze a dataset to find the “story”
* Create a Jupyter notebook for exploratory data analysis (EDA)
* Create visualization(s) using Tableau
* Articulate findings in an one-page summary/email for your cross-functional team

# Relevant Interview Questions

Completing this portfolio project will empower you to respond to the following interview topics:

* 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 a 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: Planning Stage**

* What are the data columns and variables and which ones are most relevant to your deliverable?

There are 17 columns and 16 variables with data related to taxi rides. The most relevant to my deliverables are *trip\_distance, duration, and total\_amount* but it really depends of what question one wants to answer with a helpful visualization suitable for the point one wants to convey.

So for predicting the taxi cab ride fare the variables mentioned before are the most appropriate, its is important to emphasize the selection of which graph is/are the most useful, although Tableau now uses machine learning to suggest automatically the graph.

* What units are your variables in?

The units of the variables are mostly numeric in integer and float forms.

* 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 assume that the data is accurate, complete and properly recorded. Also given the level I assume it has a normal distribution and that it does not contain outliers or anomalies. However, assumptions are not enough thus why it is important to do EDA.

* Is there any missing or incomplete data?

There is no missing data according to the results from the info() function.

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

There are present some of the most common type of formats like numeric types such as integers and floating point numbers and convenience object types such as date time.

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

It is important to remember that EDA practices are not iterative and non-sequential.

I personally will start with the Discovering and Joining process, then validating.

**PACE: Analyzing Stage**

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

Again it is important to remember that EDA practices are not iterative and non-sequential.

Because of the varying nature of datasets, the approach to exploring that data will be different each time. That means that I will need to use my logic and experience throughout the EDA process to determine which of the six practices to utilize, how many times to apply them, and when in the process you should apply them.

* 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, the practice of “joining” was performed immediately following the practice of “discovering.”

You structure the data in different time periods and segments to understand trends. In this case, perhaps would be necessary to use filtering for hiding outliers.

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

I initially assume that they are familiar for anyone with little or no technical background in my audience and are accessible and easy to read. Also that these align with the audience domain knowledge and are effective to communicate the intended message.

**PACE: Constructing Stage**

* What data visualizations, machine learning algorithms, or other data outputs will need to be built in order to complete the project goals?

One of the data visualizations that can be more needed are box-plots and histograms.

In terms of ML algorithms, BigQuery Machine Learning offers several for model prediction from the already seem linear and logistic regression models which are probably enough to achieve this project goals. It has also more complex ones such as K-means clustering, Matrix factorization and Deep NN etc.

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

I might use the discovering, structuring or cleaning processes.

* 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)?

There are 49 numbers that do not represent a drop-off location.

To eliminate the spaces in the historgram that these missing numbers would create, sort the unique drop-off location values, then convert them to strings. This will make the histplot function display all bars directly next to each other.

**PACE: Execute Stage**

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

- There are some anomalies, namely trips that have a total cost entered, but a total distance of “0.

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

- I would recommend removing outliers with a total distance recorded of 0.

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

- are there any other data points that can be an obstacle for accurate ride prediction.

- which variable has the most impact on trip fares?

- what variables are the most relevant for running regression, statistical analysis and parameter tuning?

* How might you share these visualizations with different audiences?

- Share them with the organizations directors.

- Tailoring the report to different segment of the audience: for example those who are data savvy we can use more complex terms and visualizations and those who are not we can use simple plain terms and visualizations.

Further, remember that the assistant director at NYC TLC is a person with visual impairments so by providing robust descriptions of the data. I can use things like all text descriptive text or captioned recording of the data so that my audience can explore the data themselves.