**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?

0 Unnamed: 0 22699 non-null int64

1 VendorID 22699 non-null int64

2 tpep\_pickup\_datetime 22699 non-null object

3 tpep\_dropoff\_datetime 22699 non-null object

4 passenger\_count 22699 non-null int64

5 trip\_distance 22699 non-null float64

6 RatecodeID 22699 non-null int64

7 store\_and\_fwd\_flag 22699 non-null object

8 PULocationID 22699 non-null int64

9 DOLocationID 22699 non-null int64

10 payment\_type 22699 non-null int64

11 fare\_amount 22699 non-null float64

12 extra 22699 non-null float64

13 mta\_tax 22699 non-null float64

14 tip\_amount 22699 non-null float64

15 tolls\_amount 22699 non-null float64

16 improvement\_surcharge 22699 non-null float64

17 total\_amount 22699 non-null float64

Most relevant columns:

tpep\_pickup\_datetime

tpep\_dropoff\_datetime

passenger\_count

trip\_distance

fare\_amount

extra

mta\_tax

tip\_amount

tolls\_amount

improvement\_surcharge

total\_amount

* What units are your variables in?

Refer to data dictionary file

* 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?

Handling of outliers (delete, reassign, keep them)

* Is there any missing or incomplete data?

Data for trips with total costs that have a total distiance of 0

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

Refer to jupyter notebook for details, there are string objects, int64 and float64 data types indicating that the date columns need to be change to datetime to calculate e.g. trip duration

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

- Use python function mean, median to detect range of the data values

- Calculate outliers using boxplot, as well distribution using histograms

**PACE: Analyze Stage**

* What steps need to be taken to perform EDA in the most effective way to achieve the project goal?
* Data exploration and cleaning
* Verify dimensions and measures if these are correct to know which columns/fields to use for the analyses
* Select visualization types to analyze data range
* 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.?

As there is no missing data there is no need to join additional data to this point

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

As there is the data analyst team interested in the EDA I would refer to visualizations like box plot, scatter plot, 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?

Box plots and histograms 🡪 to identify outliers

Bar plots 🡪 to identify patterns in data using certain field like total ride count per day/month, or total revenue by weekday

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

Data needs to be assessed (cleaned if there is a reasonable assumption to incorrect data representing outliers as such)

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

These are the variables having the largest impact on trip fares (trip duration, number of passengers etc)

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

Exclude outliers with total distance of 0 as this data makes no sense unless the data provider proves the opposite to this.

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

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

As a result of the conducted exploratory data analysis, the Automatidata data team considered trip distance and total amount as key variables to depict a taxi cab ride. The provided scatter plot shows the relationship between the two variables. This scatter plot was created in Tableau to enhance the provided visualization.

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

Ensuring with New York City TLC that the sample provided is an accurate reflection of their data as a whole.

* Given what you know about the data and the visualizations you were using, what other questions could you research for the team?
* Determine the variables that have the largest impact on trip fares
* Determine any unusual data points that could pose a problem for future analysis in predicting trip fares (e,g, locations that have longer durations)
* How might you share these visualizations with different audiences?

Adapt (annotate, explain in more detail) as per audience expected to attend