**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 StageA white paper with a checklist on it

AI-generated content may be incorrect.**

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

The dataset contains various columns detailing taxi trips, but for this project's goals of EDA, visualization, and creating an executive summary focused on ridership and fare factors, the most relevant variables include tpep\_pickup\_datetime and tpep\_dropoff\_datetime (essential for deriving trip duration and analyzing time trends), passenger\_count, trip\_distance, PULocationID, DOLocationID (for location analysis/mapping), RatecodeID, payment\_type, fare\_amount, tip\_amount, tolls\_amount, and total\_amount. These columns provide the necessary data for calculating key metrics, understanding trip characteristics, analyzing financial aspects, exploring temporal and spatial patterns, and ultimately addressing the core objectives, while other columns like store\_and\_fwd\_flag are less pertinent to these specific deliverables.

* What units are your variables in?

Based on the data dictionary, trip\_distance is measured in miles, while monetary values like fare\_amount, tip\_amount, tolls\_amount, and total\_amount are in US Dollars ($). The passenger\_count is a simple integer count, tpep\_pickup\_datetime and tpep\_dropoff\_datetime are date/time entries, and identifiers such as RatecodeID, payment\_type, PULocationID, and DOLocationID function as unitless categorical codes. If a duration variable is calculated, its unit (like minutes or seconds) would be determined by the calculation method used.

* 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?

My initial presumptions are that key variables like trip\_distance, calculated duration, and fare\_amount/total\_amount will be positively correlated, with distance likely being a primary driver of fare. I expect to see distinct time-based patterns, with higher trip volumes and potentially different fare dynamics during typical rush hours or weekends compared to off-peak times or weekdays. Geographically, trips are likely concentrated in certain high-traffic zones like Manhattan business districts or airports. It's also presumed that tip\_amount will correlate strongly with credit card payments (payment\_type=1) and that many key numerical variables (distance, duration, fare, tips) will exhibit right-skewed distributions, alongside the potential presence of outliers or data anomalies (like zero-distance trips) that will need confirmation and investigation during analysis.

* Is there any missing or incomplete data?

Based on initial checks like df.info(), the dataset appears complete in the sense that there are no missing (null) values across the columns. However, further exploration revealed the presence of potentially implausible entries, such as trips logged with a 0.0 trip\_distance, which indicate a different kind of data quality issue or incompleteness that warrants investigation and careful handling during the subsequent analysis and cleaning stages.

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

Upon initial loading, the dataset columns were not in a uniform format, displaying a mix of data types including int64, float64, and object. The most significant format issue identified via df.info() was that the datetime columns (tpep\_pickup\_datetime, tpep\_dropoff\_datetime) were incorrectly typed as object (strings) instead of the required datetime64[ns] format, which necessitated conversion using pd.to\_datetime() to enable proper time-based analysis and calculations like trip duration. While other columns generally had suitable initial types, correcting the datetime format was the crucial formatting adjustment needed during data preparation.

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

To begin this project, essential EDA practices include loading the data into a pandas DataFrame, followed by initial inspection using functions like .head(), .shape, .info(), and .describe() to understand its dimensions, data types, and basic statistical properties. Key subsequent steps involve verifying and correcting data types, particularly converting datetime strings to datetime objects, checking for missing values, and performing basic feature engineering such as calculating trip duration. Finally, conducting preliminary outlier reviews using descriptive statistics and initial univariate visualizations like histograms or box plots for key numerical variables are necessary starting points for analysis.

**PACE: Analyze Stage**

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

To perform EDA effectively towards the project goal of understanding ridership and fare factors, the necessary steps include first loading, inspecting, and cleaning the data—crucially converting datetime columns and calculating trip duration. Following data preparation, systematically analyze key variables like trip\_distance, duration, and fare\_amount individually using histograms and box plots to understand their distributions and identify outliers. Subsequently, investigate relationships between pairs of important variables, particularly trip\_distance versus fare\_amount, using scatter plots and correlation analysis. It's also vital to explore temporal patterns by aggregating data by month, day, or hour and visualizing trends, as well as comparing key metrics across categorical variables like payment\_type or RatecodeID using grouped analysis and bar charts, while specifically investigating anomalies like zero-distance trips, ensuring visualizations guide the entire process towards actionable insights.

* 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.?

For the core EDA based on the provided dataset, joining additional data isn't strictly required, although merging with a taxi zone lookup table would be necessary to convert PULocationID and DOLocationID into meaningful borough or neighborhood names for deeper spatial analysis or mapping. However, significant structuring is essential: this includes feature engineering like calculating the duration column and extracting time features (month, day, hour), grouping data using groupby() for aggregations (e.g., calculating totals or averages by category or time period), sorting data (sort\_values()) to view trends or extremes, and filtering the dataset, which is particularly important for investigating and potentially handling identified outliers or anomalies such as the zero-distance trips.

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

My initial assumption is that the choice of visualization depends heavily on the audience: for internal discussions with the technically proficient Automatidata team, standard EDA plots like detailed histograms, box plots, and scatter plots generated via Python are suitable for in-depth exploration. However, for the non-technical NYC TLC stakeholders, especially considering the need for accessibility (visual impairments) and inclusion in an executive summary, the visualizations must be simpler and highly interpretable at a glance; therefore, clear bar charts showing key comparisons, straightforward line graphs illustrating trends, or perhaps a very clear geographic map (if pursuing the optional Tableau task) would be most effective, all requiring careful attention to labeling, contrast, and avoiding excessive complexity.

**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?

To meet the specific goals of this Course 3 project, the primary outputs to be constructed are data visualizations within the Python notebook – including histograms, box plots (particularly for trip duration), scatter plots (like distance vs. fare), and various bar or line charts for categorical comparisons and time series analysis – all essential for the EDA. A key visualization will also need to be prepared for the executive summary, and optionally, a Tableau map dashboard can be built. While the overall client engagement aims for a fare prediction model, constructing machine learning algorithms is outside the scope of this particular EDA-focused project; other necessary outputs include the cleaned dataset resulting from the analysis, calculated features like duration, the completed PACE strategy document, and the final executive summary report.

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

Building the necessary data visualizations requires first selecting the relevant variables from the cleaned dataset and choosing an appropriate visualization type (like histogram, box plot, scatter plot, bar chart, or line chart) suited to the specific question being explored about distributions, relationships, or trends. Often, data must be further prepared specifically for plotting, such as by calculating aggregations (e.g., mean values or counts using groupby()) or filtering subsets. The visualization is then generated using Python libraries like Matplotlib and Seaborn, mapping data columns to visual elements (axes, color, etc.), and critically, the process concludes with refining the plot by adding clear titles, axis labels, adjusting formatting for readability, and ensuring accessibility for the intended audience.

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

For constructing the visualizations central to this project's EDA goals, the most applicable variables are those describing the core trip characteristics and financial aspects: trip\_distance, the calculated duration (derived from tpep\_pickup\_datetime and tpep\_dropoff\_datetime), fare\_amount, total\_amount, and tip\_amount are fundamental for plots showing distributions and relationships. Categorical variables like passenger\_count, payment\_type, and RatecodeID are key for comparative visualizations (e.g., bar charts, grouped box plots), while the datetime columns are essential for creating time-series analyses, and PULocationID/DOLocationID are needed for any location-based plots or maps.

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

Going back to the Plan stage and reflecting on the initial data assessment (using df.info()), it was determined that this specific dataset did not contain any missing (null) values. Consequently, there was no need to formulate a plan for handling missing data through methods like imputation or deletion based on nullity, and the data preparation focus could instead be directed towards addressing other identified aspects like data type conversions, outlier investigation, and handling potentially implausible values such as zero-distance trips.

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

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

Key insights emerging from the EDA and visualizations indicate that most taxi trips are relatively short (predominantly under 5 miles) and likely inexpensive, although significant outliers exist for trip\_distance, calculated duration, and fare\_amount/total\_amount, which often show right-skewed distributions. The analysis revealed distinct temporal patterns, with variations in ride counts and revenue across different months and days of the week, and potentially differences in metrics when grouped by categories like passenger\_count or VendorID. Location analysis also suggested variability, for instance, in mean trip distance based on drop-off zones. Importantly, while no missing values were found, the presence of outliers and anomalies like zero-distance trips highlights areas needing further cleaning or investigation before any modeling, while the confirmation that duration can be calculated from timestamps provides a valuable potential feature.

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

Based on the visualizations and exploratory analysis, key recommendations include urging the NYC TLC to investigate the significant number of trips recorded with zero distance to understand their cause (e.g., data errors, specific events) and improve overall data quality for future modeling. Additionally, the presence of substantial outliers in fare\_amount, trip\_distance, and calculated duration suggests a need to review these extreme cases for potential errors or fraud, while also informing Automatidata on necessary data cleaning or transformation steps before building the prediction model. Finally, the clear temporal patterns (daily, monthly) and location-based variations observed strongly suggest that time-based features and location IDs will be critical predictors in the fare model, and further analysis might offer TLC insights for operational adjustments or resource planning.

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

Based on the initial exploration and visualizations, further research questions for the team could involve a deeper dive into the identified anomalies, specifically investigating the nature and prevalence of zero-distance trips and other significant outliers in fare, distance, and duration to understand their root causes. Joining the location IDs with a taxi zone lookup table would enable valuable spatial analysis, researching how trip patterns, fares, and durations differ across specific NYC boroughs or neighborhoods. Additionally, exploring the interplay between trip distance and duration at different times of day could yield insights into traffic impacts, and further dissecting the fare components by examining the influence of different RatecodeIDs or the role of tolls\_amount on specific routes could uncover more nuances in the fare structure.

* How might you share these visualizations with different audiences?

Sharing these visualizations effectively requires tailoring the approach based on the audience: for technical colleagues within Automatidata, the entire Jupyter Notebook containing all detailed plots (histograms, box plots, scatter plots, etc.) along with the code and analytical commentary can be shared directly for a comprehensive understanding. However, when communicating with the non-technical NYC TLC team, particularly for the executive summary, only select, key visualizations that clearly convey high-level insights (like polished bar charts, line graphs showing trends, or the optional Tableau map) should be used, embedded within the summary document, and accompanied by concise, non-technical explanations, ensuring strict adherence to accessibility guidelines like high contrast and clear labeling for all stakeholders.