**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?

1. **ID**: Trip identification number
2. **VendorID**: Code indicating the TPEP provider (1= Creative Mobile Technologies, LLC; 2= VeriFone Inc.)
3. **tpep\_pickup\_datetime**: Date and time when the meter was engaged
4. **tpep\_dropoff\_datetime**: Date and time when the meter was disengaged
5. **Passenger\_count**: Number of passengers in the vehicle (driver-entered value)
6. **Trip\_distance**: Elapsed trip distance in miles reported by the taximeter
7. **PULocationID**: TLC Taxi Zone in which the taximeter was engaged
8. **DOLocationID**: TLC Taxi Zone in which the taximeter was disengaged
9. **RateCodeID**: Final rate code in effect at the end of the trip (1= Standard rate, 2=JFK, 3=Newark, 4=Nassau or Westchester, 5=Negotiated fare, 6=Group ride)
10. **Store\_and\_fwd\_flag**: Indicates if the trip record was held in vehicle memory before being sent to the vendor (Y= store and forward trip, N= not a store and forward trip)
11. **Payment\_type**: Numeric code signifying how the passenger paid for the trip (1= Credit card, 2= Cash, 3= No charge, 4= Dispute, 5= Unknown, 6= Voided trip)
12. **Fare\_amount**: Time-and-distance fare calculated by the meter
13. **Extra**: Miscellaneous extras and surcharges
14. **MTA\_tax**: $0.50 MTA tax based on the metered rate in use
15. **Improvement\_surcharge**: $0.30 improvement surcharge assessed trips at the flag drop
16. **Tip\_amount**: Tip amount (automatically populated for credit card tips; cash tips are not included)
17. **Tolls\_amount**: Total amount of all tolls paid in the trip
18. **Total\_amount**: Total amount charged to passengers (excluding cash tips)

### Most Relevant Columns for Deliverable

The columns most relevant to your deliverable, depending on the specific goals of your data analysis, might include:

1. **ID**: For unique identification of each trip.
2. **tpep\_pickup\_datetime** and **tpep\_dropoff\_datetime**: For analyzing trip duration and peak hours.
3. **Passenger\_count**: For understanding occupancy trends.
4. **Trip\_distance**: For assessing the range and extent of trips.
5. **PULocationID** and **DOLocationID**: For spatial analysis of pickup and drop-off locations.
6. **RateCodeID**: For understanding fare structure and rate application.
7. **Fare\_amount**, **Extra**, **MTA\_tax**, **Improvement\_surcharge**, **Tip\_amount**, **Tolls\_amount**, and **Total\_amount**: For comprehensive fare and surcharge analysis.
8. **Payment\_type**: For understanding payment preferences.

The relevance of each column depends on your specific analysis objectives, such as analyzing traffic congestion, fare patterns, spatial trends, or passenger behavior.

### Analyzing Traffic Congestion:

1. **tpep\_pickup\_datetime**: To identify peak times of congestion.
2. **tpep\_dropoff\_datetime**: To calculate trip duration and correlate with congestion periods.
3. **Trip\_distance**: To analyze trip lengths and their correlation with congestion.
4. **PULocationID**: To identify areas with high pickup frequencies during congested times.
5. **DOLocationID**: To identify common drop-off points in congested areas.

### Analyzing Fare Patterns:

1. **Fare\_amount**: To understand the base fare patterns across different trips.
2. **Extra**: To analyze the impact of surcharges on fare.
3. **MTA\_tax**: To account for the fixed tax in fare calculations.
4. **Improvement\_surcharge**: To see the effect of additional charges on total fare.
5. **Tip\_amount**: To analyze tipping patterns.
6. **Tolls\_amount**: To include tolls in fare analysis.
7. **Total\_amount**: To understand the overall cost to passengers.

### Analyzing Spatial Trends:

1. **PULocationID**: To determine high-demand pickup zones.
2. **DOLocationID**: To determine popular drop-off zones.
3. **Trip\_distance**: To analyze the relationship between trip distance and pickup/drop-off locations.
4. **RateCodeID**: To identify spatial patterns related to specific rate codes (e.g., trips to airports).

### Analyzing Passenger Behavior:

1. **Passenger\_count**: To understand how many passengers typically share a ride.
2. **Payment\_type**: To analyze preferred payment methods among passengers.
3. **VendorID**: To see if certain vendors are more popular or perform differently.
4. **Store\_and\_fwd\_flag**: To understand how often trips are stored before forwarding and if it affects passenger behavior.

### Overall Data Analysis Plan:

To create a comprehensive analysis plan, you could follow these steps:

1. **Data Cleaning**:
   * Ensure all datetime fields are in the correct format.
   * Handle missing or erroneous values in critical columns like Trip\_distance, Fare\_amount, etc.
2. **Exploratory Data Analysis (EDA)**:
   * Plot histograms and bar charts for columns like Passenger\_count, Payment\_type, RateCodeID.
   * Analyze time series data for pickup and drop-off times to identify peak periods.
   * Create heatmaps for PULocationID and DOLocationID to visualize spatial trends.
3. **Traffic Congestion Analysis**:
   * Calculate trip durations and plot them against different times of the day.
   * Identify high congestion zones using PULocationID and DOLocationID.
4. **Fare Pattern Analysis**:
   * Analyze the distribution of Fare\_amount, Extra, and Total\_amount.
   * Compare fare amounts across different RateCodeID and Payment\_type categories.
5. **Spatial Trends Analysis**:
   * Map out pickup and drop-off locations to visualize high-demand areas.
   * Analyze trip distances in relation to different locations.
6. **Passenger Behavior Analysis**:
   * Study the distribution of Passenger\_count.
   * Analyze payment preferences using Payment\_type data

* What units are your variables in?
*  **ID**: No units (unique identifier)
*  **VendorID**: No units (categorical code)
*  **tpep\_pickup\_datetime**: Date and time (timestamp)
*  **tpep\_dropoff\_datetime**: Date and time (timestamp)
*  **Passenger\_count**: Number of passengers (integer)
*  **Trip\_distance**: Miles (distance)
*  **PULocationID**: No units (categorical code representing TLC Taxi Zone)
*  **DOLocationID**: No units (categorical code representing TLC Taxi Zone)
*  **RateCodeID**: No units (categorical code)
*  **Store\_and\_fwd\_flag**: No units (categorical flag: Y/N)
*  **Payment\_type**: No units (categorical code)
*  **Fare\_amount**: US Dollars (currency)
*  **Extra**: US Dollars (currency)
*  **MTA\_tax**: US Dollars (currency)
*  **Improvement\_surcharge**: US Dollars (currency)
*  **Tip\_amount**: US Dollars (currency)
*  **Tolls\_amount**: US Dollars (currency)
*  **Total\_amount**: US Dollars (currency)
* 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?

### Traffic Congestion

1. **Peak Hours**: Higher trip frequency and longer trip durations during rush hours (e.g., morning and evening commutes).
2. **High Demand Zones**: Certain locations (e.g., downtown, airports) will show higher pickup and drop-off frequencies, especially during peak hours.
3. **Trip Durations**: Trip durations might be longer during peak hours due to increased traffic congestion.

### Fare Patterns

1. **Fare Variability**: Variability in fare amounts due to different RateCodeID, with higher fares for special rates like JFK, Newark, and negotiated fares.
2. **Surcharges**: Surcharges (Extra, MTA\_tax, Improvement\_surcharge) will significantly affect the total fare.
3. **Tips**: Higher tips for credit card payments compared to cash payments.

### Spatial Trends

1. **Popular Locations**: Some zones will have consistently high pickup and drop-off counts, potentially indicating popular areas or transportation hubs.
2. **Trip Distance**: Trips to and from airports or out-of-city zones (JFK, Newark) will have longer distances compared to intra-city trips.

### Passenger Behavior

1. **Passenger Count**: Most trips will likely have a low passenger count (1-2 passengers).
2. **Payment Preferences**: Credit card payments might be more common than cash payments.
3. **Store and Forward Trips**: A minority of trips will be stored and forwarded (Store\_and\_fwd\_flag = Y).

### Data Quality and Integrity

1. **Missing Data**: Some records might have missing or incorrect values, especially in fields like Passenger\_count, Tip\_amount, and Trip\_distance.
2. **Outliers**: There might be outliers in terms of extremely high or low fare amounts, trip distances, or passenger counts that need to be addressed.

### Confirming or Denying Presumptions

During the EDA, these initial presumptions will need to be confirmed or denied by:

* **Visualizations**: Using histograms, bar charts, and scatter plots to observe distributions and relationships.
* **Statistical Analysis**: Calculating summary statistics (mean, median, mode, standard deviation) to understand central tendencies and variability.
* **Correlation Analysis**: Checking for correlations between variables such as trip distance and fare amount, pickup time and trip duration, etc.
* **Geospatial Analysis**: Mapping pickup and drop-off locations to identify spatial patterns and high-demand zones.
* **Time Series Analysis**: Analyzing trip data over time to identify trends and patterns related to peak hours and days of the week.
* Is there any missing or incomplete data?

None

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

1. **Data Types**: Ensuring all columns are in the correct data types.
2. **Consistency**: Ensuring consistency within each column (e.g., no mixed data types within a column).
3. **Formatting**: Ensuring date and time formats are consistent.
4. **Units**: Ensuring numerical values are in the same units.

Given the data types provided earlier, let's evaluate each column:

1. **Unnamed: 0**:
   * Data Type: int64
   * **Evaluation**: This column appears to be an index and can be dropped if unnecessary.
2. **VendorID**:
   * Data Type: int64
   * **Evaluation**: Consistent, categorical numeric codes.
3. **tpep\_pickup\_datetime**:
   * Data Type: object (should be datetime)
   * **Evaluation**: Needs conversion to datetime to ensure consistency.
4. **tpep\_dropoff\_datetime**:
   * Data Type: object (should be datetime)
   * **Evaluation**: Needs conversion to datetime to ensure consistency.
5. **passenger\_count**:
   * Data Type: int64
   * **Evaluation**: Consistent integer values.
6. **trip\_distance**:
   * Data Type: float64
   * **Evaluation**: Consistent floating-point values representing miles.
7. **RatecodeID**:
   * Data Type: int64
   * **Evaluation**: Consistent, categorical numeric codes.
8. **store\_and\_fwd\_flag**:
   * Data Type: object (should be category)
   * **Evaluation**: Needs conversion to a categorical type.
9. **PULocationID**:
   * Data Type: int64
   * **Evaluation**: Consistent, categorical numeric codes.
10. **DOLocationID**:
    * Data Type: int64
    * **Evaluation**: Consistent, categorical numeric codes.
11. **payment\_type**:
    * Data Type: int64
    * **Evaluation**: Consistent, categorical numeric codes.
12. **fare\_amount**:
    * Data Type: float64
    * **Evaluation**: Consistent floating-point values representing dollars.
13. **extra**:
    * Data Type: float64
    * **Evaluation**: Consistent floating-point values representing dollars.
14. **mta\_tax**:
    * Data Type: float64
    * **Evaluation**: Consistent floating-point values representing dollars.
15. **tip\_amount**:
    * Data Type: float64
    * **Evaluation**: Consistent floating-point values representing dollars.
16. **tolls\_amount**:
    * Data Type: float64
    * **Evaluation**: Consistent floating-point values representing dollars.
17. **improvement\_surcharge**:
    * Data Type: float64
    * **Evaluation**: Consistent floating-point values representing dollars.
18. **total\_amount**:
    * Data Type: float64
    * **Evaluation**: Consistent floating-point values representing dollars.

### Summary

* **Date and Time Columns**: tpep\_pickup\_datetime and tpep\_dropoff\_datetime need conversion to datetime for consistency.
* **Categorical Columns**: store\_and\_fwd\_flag should be converted to a categorical type.
* **Numerical Columns**: Consistent in terms of data types and units.
* Which EDA practices will be required to begin this project?

To begin EDA, clean and convert data types, check for missing values, generate summary statistics, analyze time-based patterns, and examine spatial trends.

**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, start by cleaning the dataset, including handling missing values and converting data types. Next, generate summary statistics and visualize distributions to understand data characteristics. Analyze time-based patterns and spatial trends to uncover insights relevant to traffic congestion, fare patterns, and passenger behavior. Finally, identify and address outliers to ensure robust results.
* 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.?

You may need to join additional data if external datasets, such as weather or events, could enhance the analysis. Structuring the dataset involves filtering relevant time frames or locations, sorting data by key variables like trip distance or fare amount, and aggregating data to identify patterns or trends. Ensure data is cleaned and transformed consistently to facilitate accurate analysis.

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

For the intended audience, clear and informative visualizations such as heatmaps for spatial trends, time series plots for peak hours, histograms for fare distributions, and scatter plots for trip distances vs. fares will be effective. These visualizations can convey patterns and insights intuitively, helping to highlight key findings and trends in traffic congestion, fare patterns, and passenger behavior.

**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 complete the project goals, you will need to create visualizations such as heatmaps for spatial patterns, line graphs for time-based trends, and scatter plots for relationships between variables. Machine learning algorithms like clustering (e.g., K-means) for identifying traffic patterns, regression analysis for fare predictions, and classification models for predicting payment types may be useful. Additionally, generating summary statistics and correlation matrices will aid in understanding data relationships and trends.

* What processes need to be performed in order to build the necessary data visualizations?
*  **Data Preparation**: Clean and structure the data, including handling missing values, converting data types, and filtering relevant subsets.
*  **Exploratory Analysis**: Conduct initial analysis to understand data distributions, relationships, and patterns. Generate summary statistics and preliminary plots.
*  **Visualization Design**: Choose appropriate types of visualizations (e.g., heatmaps, scatter plots, time series) based on the data and analysis goals.
*  **Plotting**: Use visualization libraries (e.g., Matplotlib, Seaborn, Plotly) to create the plots. Customize visuals for clarity, including axis labels, titles, and legends.
*  **Interpretation**: Analyze the visualizations to extract insights, identify trends, and highlight key findings.
*  **Iteration**: Refine visualizations based on feedback and additional analysis to ensure they effectively communicate the desired insights.
* Which variables are most applicable for the visualizations in this data project?

 **tpep\_pickup\_datetime and tpep\_dropoff\_datetime**:

* **Use**: Time series analysis to identify peak travel times and patterns.
* **Visualizations**: Time series plots, heatmaps of pickup and drop-off times.

 **PULocationID and DOLocationID**:

* **Use**: Spatial analysis to understand pickup and drop-off locations.
* **Visualizations**: Heatmaps, geographic maps, and cluster maps.

 **trip\_distance**:

* **Use**: Analyzing the distribution of trip distances.
* **Visualizations**: Histograms, scatter plots (e.g., distance vs. fare).

 **fare\_amount, extra, mta\_tax, tip\_amount, tolls\_amount, improvement\_surcharge, and total\_amount**:

* **Use**: Examining fare and cost components.
* **Visualizations**: Box plots, histograms, and pie charts of fare distributions.

 **passenger\_count**:

* **Use**: Analyzing the number of passengers per trip.
* **Visualizations**: Histograms, bar charts.

 **RatecodeID and payment\_type**:

* **Use**: Understanding fare types and payment methods.
* **Visualizations**: Bar charts, pie charts.

 **store\_and\_fwd\_flag**:

* **Use**: Analyzing the impact of data storage issues.
* **Visualizations**: Bar charts or pie charts showing the proportion of store-and-forward trips.
* Going back to the Plan stage, how do you plan to deal with the missing data (if any)?
* **Identification**:
  + **Assess Missing Data**: Identify columns with missing values and the proportion of missing data.

python

Copy code

missing\_values = df.isnull().sum()

print(missing\_values)

* **Analysis**:
  + **Determine Patterns**: Analyze if there is a pattern or reason behind the missing data. Check if missing values are random or correlated with other variables.
* **Handling Strategies**:
  + **Removal**: If missing data is minimal and random, remove rows with missing values.

python

Copy code

df = df.dropna()

* + **Imputation**:
    - **Numerical Columns**: Impute missing values with the mean or median of the column.

python

Copy code

df['trip\_distance'].fillna(df['trip\_distance'].median(), inplace=True)

* + - **Categorical Columns**: Impute missing values with the mode of the column.

python

Copy code

df['RatecodeID'].fillna(df['RatecodeID'].mode()[0], inplace=True)

* + **Flag and Analyze**: Create a new binary column indicating missing values to analyze if the missing data itself reveals any patterns.

python

Copy code

df['missing\_trip\_distance'] = df['trip\_distance'].isnull().astype(int)

* **Documentation**:
  + **Record Decisions**: Document the methods used for handling missing data and the rationale behind them.

By following these steps, you ensure that missing data is handled systematically and does not bias the analysis.

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

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

 **Temporal Patterns**:

* **Peak Travel Times**: Identifying peak hours or days for pickups and drop-offs, revealing busy periods.
* **Seasonal Trends**: Detecting any seasonal variations or trends over different months or times of the year.

 **Spatial Patterns**:

* **Hotspots**: Pinpointing high-traffic zones for pickups and drop-offs, showing popular locations.
* **Geographic Distribution**: Understanding spatial distribution of trips to optimize routing and resource allocation.

 **Fare and Cost Analysis**:

* **Fare Distribution**: Analyzing fare amounts and identifying typical fare ranges and outliers.
* **Cost Components**: Evaluating the impact of additional charges (e.g., extra fees, taxes) on total fare.

 **Passenger Behavior**:

* **Passenger Counts**: Revealing the average number of passengers per trip and identifying any patterns.
* **Payment Methods**: Understanding the distribution of payment types and identifying any trends.

 **Data Quality Issues**:

* **Missing Data**: Identifying patterns in missing data and their potential impact on the analysis.
* **Outliers**: Detecting unusual data points that may require further investigation or removal.
* What business and/or organizational recommendations do you propose based on the visualization(s) built?
*  **Optimize Resource Allocation**: Focus on high-traffic pickup and drop-off zones to allocate more vehicles during peak hours and improve service efficiency.
*  **Adjust Pricing Strategies**: Review fare distributions and additional charges to ensure competitive pricing and consider dynamic pricing based on time or location.
*  **Improve Operational Efficiency**: Use time-based insights to reduce wait times and optimize routes, enhancing overall operational efficiency and customer satisfaction.
*  **Address Data Quality**: Investigate any patterns in missing data or outliers to improve data collection and processing methods for more accurate future analyses.
* Given what you know about the data and the visualizations you were using, what other questions could you research for the team?
*  **Trip Duration Analysis**: How do trip durations vary by time of day or location, and what factors contribute to longer or shorter trips?
*  **Impact of Weather**: How do weather conditions affect trip distances, fare amounts, or pickup and drop-off locations?
*  **Driver Behavior**: Are there any noticeable patterns or anomalies in driver behavior based on trip data, such as frequent detours or route choices?
*  **Fare Prediction**: Can we build a predictive model to estimate fare amounts based on trip distance, time, and other factors?
*  **Passenger Trends**: How do passenger counts correlate with time of day, location, or fare amount, and are there trends in group travel?
*  **Payment Method Analysis**: Are there differences in fare amounts or trip patterns based on the payment method used (e.g., credit card vs. cash)?
* How might you share these visualizations with different audiences?

 **Executive Team**:

* **Format**: High-level dashboards or summary reports.
* **Focus**: Key insights, trends, and actionable recommendations.
* **Tools**: PowerPoint presentations, PDF reports, or interactive dashboards (e.g., Tableau, Power BI).

 **Operational Staff**:

* **Format**: Detailed visualizations with operational metrics.
* **Focus**: Practical insights for daily operations, such as peak hours and high-traffic areas.
* **Tools**: Interactive dashboards, detailed charts in reports.

 **Data Analysts/Scientists**:

* **Format**: Comprehensive visualizations and exploratory plots.
* **Focus**: Data patterns, correlations, and statistical analyses.
* **Tools**: Jupyter notebooks, data visualization tools (e.g., Matplotlib, Seaborn), and shared data files.

 **Customers/Clients**:

* **Format**: Simplified, user-friendly visualizations.
* **Focus**: Insights relevant to customer experience, such as fare trends or service areas.
* **Tools**: Infographics, summary dashboards, or interactive web visualizations.