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

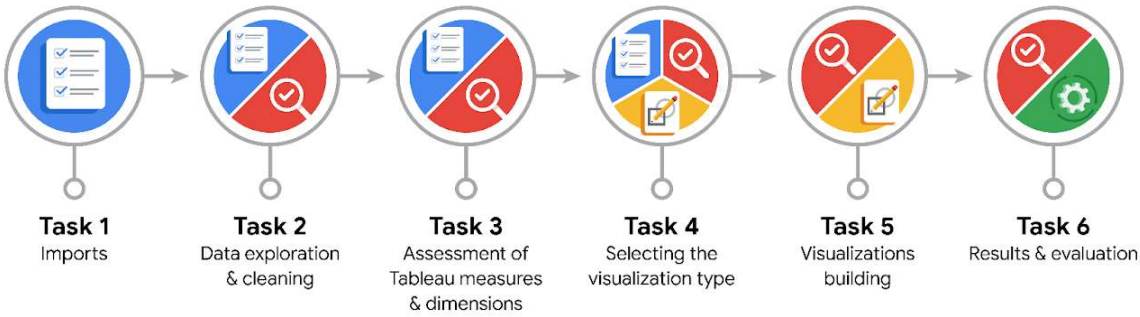
Relevant Interview Questions

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?

Among the most useful variables for this analysis were trip distance, total amount, tip amount, passenger count, pickup datetime and dropoff datetime.

- What units are your variables in?

tpep_pickup_datetime and tpep_dropoff_datetime were converted to datetime. Passenger_count is an int64, and the remaining variables are float64.

- 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 noticed that there are a few data points that appear to be outliers, which could be due to incorrect entry, collection, or other reasons. However, it's important to investigate these outliers to determine if they are genuine or simply errors.

- Is there any missing or incomplete data?

There are no missing data columns, but outliers must be investigated before model development.

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

The dataset consists of various formats including float64(8), int64(7), and object(3).

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

1. Use functions like head(), tail(), describe(), and info() to view data structure. 2. It is crucial to identify outliers, missing values, and erroneous entries in a dataset. Visualizing variable distributions with plots can help detect anomalies, which should be followed up by correcting the issues with the source. 3. Ensuring the data is in the correct format is also key, e.g.; dates entered as a string being converted to datetime for easier manipulation and analysis.

PACE: Analyze Stage

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

1. Perform a preliminary analysis of the data to resolve missing values, outliers, anomalies and to understand the distribution of key variables.
2. Visualize key variables with plots to better understand these distributions, outliers and anomalies.
3. Once these anomalies are identified, decide on the best approach to address 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.?

- Based on our analysis, no additional data needs to be added. However, to answer some additional questions, we may require more specific trip location information.
- To make the analysis easier, the tpep_pickup_datetime and tpep_dropoff_datetime columns were converted to datetime. Month and Day columns were created to calculate the total number of rides per month and day. They were sorted by weekdays and months to ensure that they are in the correct order for plotting.

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

- Visualizing the distribution of variables, outliers, and other anomalies can be effectively achieved through Bar Plots, Histograms, and Box Plots. The choice of which plot to use depends on the level of knowledge of the viewers. Additionally, a Scatter Plot is an ideal choice to visualize the correlation or relationship between two variables, like Trip Distance and Total Fare Amount.

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 get a better understanding of the distribution of important variables, we can use Box Plots and Histograms. Additionally, Bar Plots can be employed to depict the relationship between these variables. • Statistical tests can be used to determine the significance of any perceived relationships, such as T-Tests for a mean, independent or paired means.

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

• As part of the Exploratory Data Analysis (EDA) process, it is essential to clean and format the dataset to avoid skewed or inaccurate visualizations. • Once the dataset is cleaned, visualizations of key variables can be produced using libraries like SeaBorn or matplotlib.pyplot.

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

tpep_pickup_datetime and tpep_dropoff_datetime, passenger_count, trip_distance, payment_type, VendorID, fare_amount and tip_amount.

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

• To begin with, you can use functions like info() and describe() to identify the columns and rows that contain missing data. After identifying them, you need to assess how significant these missing values are to your overall analysis. • Based on this, you can take appropriate measures like removing rows with missing values, imputing the means or median, or contacting the data source to obtain the missing information, if possible.

PACE: Execute Stage

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

  - There is significant right-skew in many of the key variables. Substantial outliers are present in total_amount, tip_amount, passenger_count and trip_distance that require further investigation. These may be errors, but they may also be legitimate.

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

  - Considering the lower volume of rides during February, July, August, and September, it may be financially beneficial to reduce the number of cabs deployed to save costs. However, to make this proposal more realistic, we need more specific data such as the locations of pick-ups and drop-offs.
  - This strategy could also be used on a daily basis, especially on Mondays and Sundays, which have lower ride volume than other days of the week. The histogram of rides by drop-off location shows that certain areas experience higher volume than others. Thus, we can utilize this information to reduce the deployment of cabs in low-volume areas to save costs and potentially increase volume in popular locations to improve service.

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

  - Can we use the data we have to lower costs during periods of lower volume?
  - Additionally, is the company losing money for different amounts to passengers based on total_amount and trip_distance? To accurately answer this question, we need to consider the cost of trips, including fuel expenses.

- How could you share these visualizations with different audiences?

  - The Data Analysis team can use more detailed plots and summary statistics. Those with less knowledge and time can look at simplified plots that only show key variables and the significance of the overall business initiative.