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

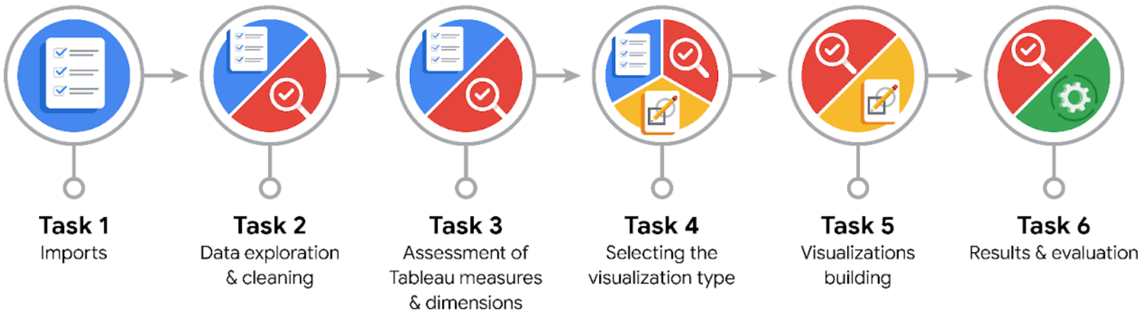
Relevant Interview Questions

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.


Task 1
Imports
Task 2
Data exploration & cleaning
Task 3
Assessment of Tableau measures & dimensions
Task 4
Selecting the visualization type
Task 5
Visualizations building
Task 6
Results & evaluation

Data Project Questions & Considerations

PACE: Plan Stage

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

The data columns include the following:

['trip_id', 'VendorID', 'tpep_pickup_datetime', 'tpep_dropoff_datetime', 'passenger_count', 'trip_distance', 'RatecodeID', 'store_and_fwd_flag', 'PULocationID', 'DOLocationID', 'payment_type', 'fare_amount', 'extra', 'mta_tax', 'tip_amount', 'tolls_amount', 'improvement_surcharge', 'total_amount'].

For my deliverables, I consider the most relevant variables to be:

'fare_amount', 'trip_distance', 'tpep_pickup_datetime', tpep_dropoff_datetime, 'payment_type'.

The fare amount and trip distance are obvious choices. The are both continuous variables. I have included the datetime to perform some time series analysis and payment types variables to see what payment method was used most. This could help understand user payment preference.

- What units are your variables in?

  - Fare amount is in US dollars
  - Trip distance in miles

- 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 presume there is an association between trip distance and fare amount. I expect that the further the trip the higher the fare. I also presume this should be related to the difference between the pickup time and drop off time.

- Is there any missing or incomplete data?

There is no missing or incomplete data observed in the dataset.

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

The pieces (variables) are of different formats. For example, the fare amount is numeric while the time pieces are a datetime object.

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

There following EDA practices will be required:

Discovering, structuring, joining, validating, cleaning and presenting.

PACE: Analyze Stage

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

First, it is important to determine where the data was coming from and what each variable meant . I determine that for the project goal, EDA will begin with discovering practices to fully understand the client's data and plan how to use them. This will be followed by structuring the data in a consistent way and cleaning the data to get rid of anomalies and outliers on the given dataset.

- 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 this analysis it will be helpful to add a **duration** variable to the dataset. Ths is the different between drop off and pic up time measured in mins.

The following structuring should be done on this dataset:

- Sort the dataset by fare amount and trip distance.

- Filter the dataset by payment method.

- Group the data by weekday, month, hours, duration

- Filter by Fare amount and trip distance that is far from the average range.

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

Two important variables for this project are trip distance and fare amount. They are both numeric and continuous. My initial assumption is to use the Scatter plot to show the association between both.

Regarding the payment method, a bar chart would be appropriate to show this distribution of trip across payment methods.

Box plot can be used to show association between fare amounts.

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?

The following data visualization will be built in order to complete the project goals: Bar chart, boxplot, histogram, dashboard, and scatter plots. In addition, a predictive machine learning model will be built.

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

The following processes are needed for to perform effective data visualization:

Know your audience, choose the right visualizations, use color effectively and keep it simple.

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

The variables most applicable for the visualization of this project are trip distance, total amounts, fare amount, payment method, trip duration(min), tip amount, and vendor. There will also include some derived variables such as Months and Day of trip.

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

There is no missing data in the dataset provided for this project.

PACE: Execute Stage

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

The dataset have some outliers that we will need to make decisions on prior to designing a model.

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

Based on your visualization, I propose the organization investigate why ride counts for Monday and Sunday are less compared to other days of the week.

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

There are several trips that have a trip distance of 0.0. What might those trips be? Will they impact our model? There are also 33 trips with zero passengers, what might those trips be? These questions will be researched for the team.

Also the data includes a trip with negative \`duration\` in minutes. This doesn't make sense. There are several trips with zero duration in minutes yet either their \`trip_distance\` or \`fare_amount\` is greater than zero. What might those trips be?

- How might you share these visualizations with different audiences?