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

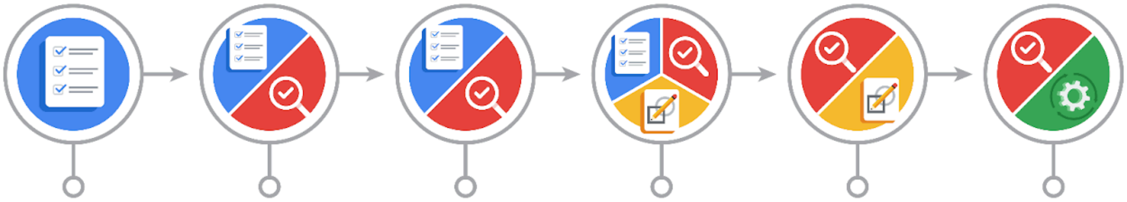
Relevant Interview Questions

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?

Vendor ID (of type 1 and 2), pickup and drop off times (tpep_pickup_datetime, tpep_dropoff_datetime), drop off location ID (DOLocationID), tip amount (Tip_amount)

- What units are your variables in?

Datetime, int64 and float64, dollar

- 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?

That data is taken with integrity. Toll amounts are exact and the drop off location is outside the city if any toll amount is recorded. Trip identification numbers are distinct. The not store and forward trip data is out of consideration for this analysis. The meter of every taxi whose trip is recorded is functioning normally and the readings are correct. Tip amount is not greater than total amount. Toll amount is not greater than total amount. Same for the toll amount, improvement surcharge, MTA_Tax, fare_amount.

- Is there any missing or incomplete data?

No.

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

There are three kinds of formats int64, object, float64

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

Check the size, shape and makeup of the dataset.

PACE: Analyze Stage

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

Use the following methods: -

df.head()

df.size

df.describe()

df.info()

- 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.?

No. Datetime objects are further broken into month and day.

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

The visualizations are easy to understand, standard chart types are being used, Key insights or the purpose of the visualisation can be easily understood.

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?

Barplot, Histogram.

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

Sorting data, size of the visualization, use python libraries such as seaborn to build the histogram. Set the bin ranges, x and y labels, plot title.

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

Trip distance, Total amount, Tip amount, Vendor, passenger_count, DOLocationID

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

Remove it as it is not to be taken into account because missing values can affect the total amount of a trip. Missing datetime info i.e. the pickup and drop off date time can affect the further breakdown of the column.

PACE: Execute Stage

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

The following are top three insights that are taken from EDA and visualizations: -

1. None of the rows have null value rather the value can be zero.

2. The highest total amount is approximately 8.5 reaching a count of 1800. No total amount is reported between 0.1 to 2.6 The outlier

3. 200 is the maximum tip amount and is an outlier.

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

The tip amount should be limited. A more integral approach at reporting trips as passenger count of 0 has reported some mean tips. Fare amounts can be raised based on insights from different seasons.

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

Seasonal variations and how they affect the number of rides.

- How might you share these visualizations with different audiences?

MS presentations or through executive summaries etc.