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:

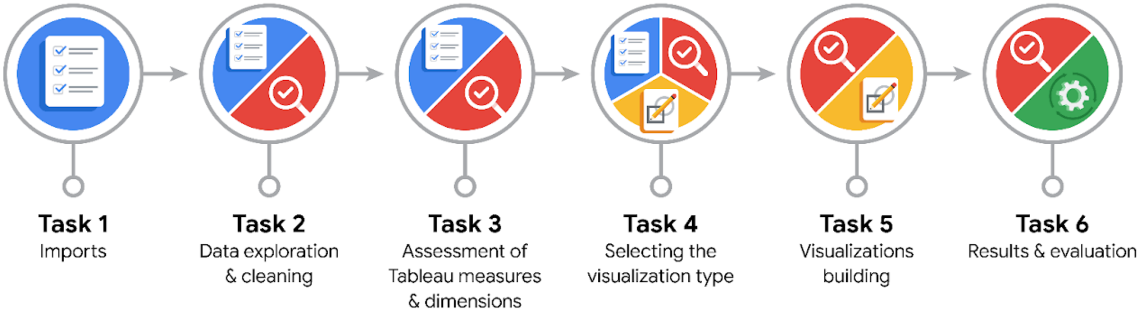
- How would you explain the difference between qualitative and quantitative data sources?
    - Qualitative data is data that is categorical and usually texual, such as days of the week, or the colour of cars
    - Quantitative data is data that is numerical and represents an aggregate amount of something, such as amount paid per purchase, total population of a city, height of a person
- Describe the difference between structured and unstructured data.
    - Structured data is organized and formatted data that can be represented in a table, with rows and columns

    - Unstructured data is data that cannot be represented in a table, and doesn't have strict organization or formatting. Unstructured data includes data like audio files and images.

- Why is it important to do exploratory data analysis?
    - EDA results in cleaned and well-formatted data, and helps us understand the structure and trends of the data we are working with.
- How would you perform EDA on a given dataset?
    - The EDA process is cyclical and heavily depends on the data we are working with. However a typical EDA might proceed as such:
        - Discovering the data types and outliers within the data
        - Joining the dataset to other datasets to supplement important information
        - Validating the dataset to check for missing values, mistakes or other errors
        - Structuring the data, separating it into categories and adding calculated columns in order to understand trends within the data, thus identifying anomalies and outliers
        - Cleaning the dataset to remove such anomalies and errors
        - Validating the dataset again to ensure that the data is free of errors
        - Presenting the analysis in the form of visualizations to stakeholders
- How do you create or alter a visualization based on different audiences?
    - Create filters that the audiences may use to alter the visualizations to their needs
- How do you avoid bias and ensure accessibility in a data visualization?
    - To avoid bias, ensure the axes of your graphs begin at 0 and use a regualr scale, else you risk inflating the differences between data points
    - For accessibility, ensure the colours you have chosen are clearly visible to people with colour vision deficiencies. Also include alt text to allow visually impaired audiences to experience the visualization
- How does data visualization inform your EDA?
    - Data visualization allows us to clearly see the trends within the dataset, identify outliers and missing data

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?

trip_distance and total_amount – the distance of the taxi cab ride and the total fare paid

- What units are your variables in?

trip_distance is in miles, total_amount is in US Dollars

- 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?

total_amount increases as trip_distance increases

- Is there any missing or incomplete data?

No, there is no missing or incomplete data in this dataset

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

No, there are some data that are floats, some integers and some strings

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

Discovering, Validating, Cleaning, Structuring and Presenting

PACE: Analyze Stage

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

Discovering, Validating, Cleaning, Validating, Structuring, Validating, Presenting

- 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.?

There is no need to add more data to this dataset. We will likely need to filter the total fare by payment type, passenger count, and time, as well as sort the dataset by trip distance and total fare.

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

We can use box plots, histograms and scatter plots to present our findings to the team. The box plots will show the spread of data, histograms to show the distribution of data over categories such as day, month, passenger count, and trip distance. The scatter plot will show any correlation between trip distance and total fare.

PACE: Construct Stage

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

Data needs to be cleaned, filtered, sorted and grouped into categories before we can build the necessary visualizations

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

trip_distance and total_amount – the distance of the taxi cab ride and the total fare paid

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

Depending on the kind of data and amount that is missing, we can consider either filling in the data with supplement datasets, filling in data based on extrapolations of the current data, or removing records with missing data

PACE: Execute Stage

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

The median trip distance and total fare is 5miles and $10, with some outliers. There is a positive correlation between trip distance and total fare. There are more rides and thus more revenue in the middle of the week compared to Mondays/Sundays, and more rides/revenue in March than in July. Lastly, there are data points with 0 distance travelled but charged a non-0 total fare

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

Offering deals and discounts on the weekend/Mondays could increase revenue. Increasing the number or drivers during busy periods such as the middle of the week or during certain months may be able to alleviate potential bottlenecks which will reduce time waiting for a taxi and also increase revenue.

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

We could investigate the correlation between time taken for each trip and the trip distance or total fare

- How might you share these visualizations with different audiences?

Present as a Tableau dashboard, allow for filters so different audiences can drill down on the different aspects that interest them