PACE Strategy Document

Introduction

This PACE strategy document can be used 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.

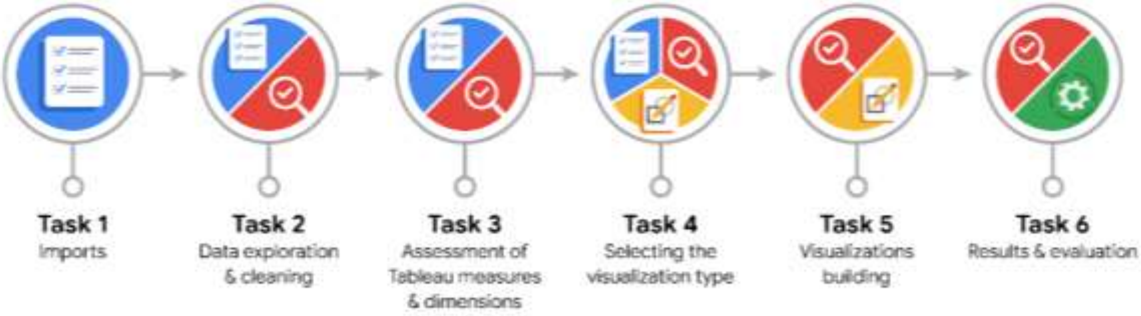
Relevant Interview Questions

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.


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?

Passenger count, trip duration, trip distance, tip amount, total amount

- What units are your variables in?

Minutes, miles, 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?

The basic presumptions include higher amount for trips with larger distances and during rush hour.

- Is there any missing or incomplete data?

Yes trip distance contains null and small values even for significant trip durations. Cash tips are not included. Also vehicle type is not included. Limo costs more than normal taxis but this info is not there in data.

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

Data types are different but format is same.

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

All of 6 EDA practices: Explore, Clean, Input Validation, Structure, Join, Visualize

PACE: Analyze Stage

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

Perform all steps properly with ethics and accessibility in consideration.

- 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.?

Further data might be needed to analyze further like weather conditions, holiday season. Also if available vehicle type data should be readily added to this dataset. Sorting by distance. Filter out null values.

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

Distance vs Total amount line chart, Box plot for total amount, Comparison of payment types bar/line, Trip duration vs Distance line chart, Passenger count vs Total amount for similar distances bar chart,

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?

Visualizations stated above. Machine learning not yet learned.

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

Data cleaning, structuring then code using python and using tableau.

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

Distance, Trip duration and Total amount

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

Give a thought to why data is missing and the drop na if no other choice.

PACE: Execute Stage

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

➢ Total amount is related to distance and time duration both.
➢ For more than 6 passengers the total amount takes a sudden spike.
➢ Some rides have a large total amount even though the trip distance is 0.
➢ The dataset contains a lot of outliers and error values like negative total amount and trip duration.
➢ Sundays record the most no. of rides and thus highest total revenue and Wednesday the least.
➢ The mean trip distance is normally distributed for the Location IDs.
➢ Also, some locations are highly popular while some are rarely visited by riders.

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

To proceed with handing outliers, eliminating errors and building models on the basis of relations found.

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

Why some rides have high unusually high total amount even though trip distance is negligible

- How might you share these visualizations with different audiences?

Different visualizations are built with labels and accessibility considerations.