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

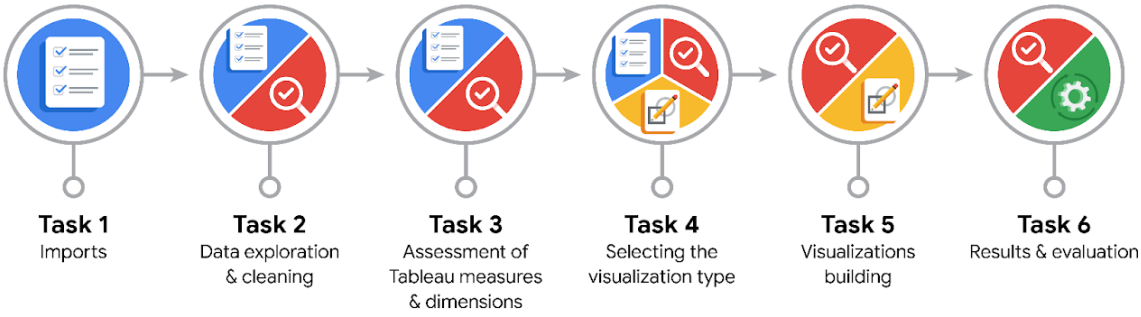
Relevant Interview Questions

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?
- 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 a 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: Planning Stage

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

  The main variables are Total Amount and Trip Distance, which corresponds to the price of the trip in dollars and the distance traveled 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?

  There are registers where there is a fee paid but there is no distance traveled. Additionally, there are prices per trip which are too expensive compared to other trips with the same distance.

- Is there any missing or incomplete data?

  Yes, there are registers with a lack of data about the distance traveled.

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

  It will be required to evaluate and possibly remove missing data and outliers.

PACE: Analyzing Stage

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

First, import the data to a Jupyter Notebook and use descriptive statistics to understand the nature of the variables and the dataset. Second, it is important to visually evaluate the distribution of the variables by creating boxplots, and its relationship through scatter plots.

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

It is better to show a scatter plot so they can see the distribution of the data and some kind of trends. Additionally, it would be helpful to indicate the values to clean so they can understand their nature.

PACE: Constructing Stage

- What data visualizations, machine learning algorithms, or other data outputs will need to be built in order to complete the project goals?

To develop the visualizations it is helpful to develop a line chart to visually understand the correlation of the data, additionally, depending on the correlation between variables it can vary the machine learning algorithms to be applied.

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

The variables can be filtered to perform a deeper analysis of the data and have a better perception of its nature in the visualizations. For this task, a bar chart can be created to observe the relation between the cost and distance. In addition, Tableau is a good option for the scatter plot, because it has an interface more specific for this kind of data and task.

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

Due to their nature it is important to understand what happened in those cases. However, it is probable that at the end of the day it would be necessary to delete that missing data because of the negative effect on the standard deviation.

PACE: Execute Stage

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

  It is recognized that the main variables were cleaned to perform the development of the predictive model. Furthermore, the visualization helped to understand the missing values and the nature of the dataset.

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

  I would recommend collecting more information about the units of the data and the precision of its measurement. In my opinion, it is essential to evaluate the margin of error of each of the main variables, so the analysis can be more precise.

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

  I would evaluate which other variables can affect the variation rate of the main variables, so this variable can be considered as a third variable to analyze and would be able to improve the accuracy of the model without causing overfitting.

- How might you share these visualizations with different audiences?

  The visualization should include information about the labels and units of the cost and duration, in addition, the visualization should represent the key features of the graph.