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

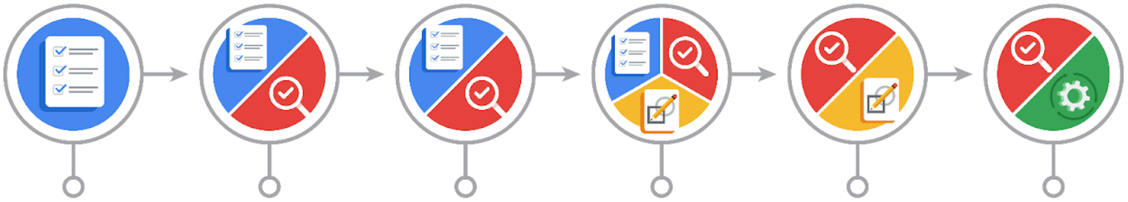
Relevant Interview Questions

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 “label” variable because it showcases the target type of user to be evaluated.

- What units are your variables in?

Mostly numeric, only label and device are objects “strings”

- 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 user device isn't related at all to user churn.

- Is there any missing or incomplete data?

Yes, label contains some missing data

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

Yes, they are fairly consistent between each column

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

Exploration of the data by using visualization and creating new columns to get new insights.

PACE: Analyze Stage

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

By using a specific structure that starts by exploring the different types of rows, and columns, datatypes and general stats, after that we can follow to create visualizations to identify any outlier or anomalies in data, and finally we can start exploring insights and discovering patterns.

- 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.?

We can create new data by creating new columns using mathematical operations like division, quartile, etc...

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

For the “object” type of data we can use pie charts or histograms that analize the distribution of devices and labels, for the numerical data we can use specifically histograms and scatter plots.

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?

We will need to create specific data visualization that showcases label distribution and also generate data that showcases insights that are relevant for the stakeholders.

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

First gather the data, and using the documentation we can understand wich parameters are needed to showcase our visualization in a clear and non biased way.

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

As I said, “label” and “device” are quite important to create a visualization of their distribution, also numeric data that can be visualized as time series with histograms

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

By first identifying the outliers and trying to comprehend its root cause, then if the data is outlier because is mislabeled it can be changed, if not we can choose to fill it manually or to “crop it” using the quartile range

PACE: Execute Stage

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

- The more times users used the app, the less likely they were to churn. While 40% of the users who didn't use the app at all last month churned, nobody who used the app 30 days churned.

-Distance driven per driving day had a positive correlation with user churn. The farther a user drove on each driving day, the more likely they were to churn.

-Number of driving days had a negative correlation with churn. Users who drove more days of the last month were less likely to churn.

- Users of all tenure

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

I don't have any particular recommendation right now, I propose to dive deeper into the relationship of the datapoints, specifically those related to “labels”

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

Who are the type of people that have an extremely high number of drives? Why do retained users have fewer drives than churned users? What is the demographic of churned users?

- How might you share these visualizations with different audiences?

By properly formatting them in a way that is easy to understand and comprehend, also having high contrast colors so its more accessible.