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

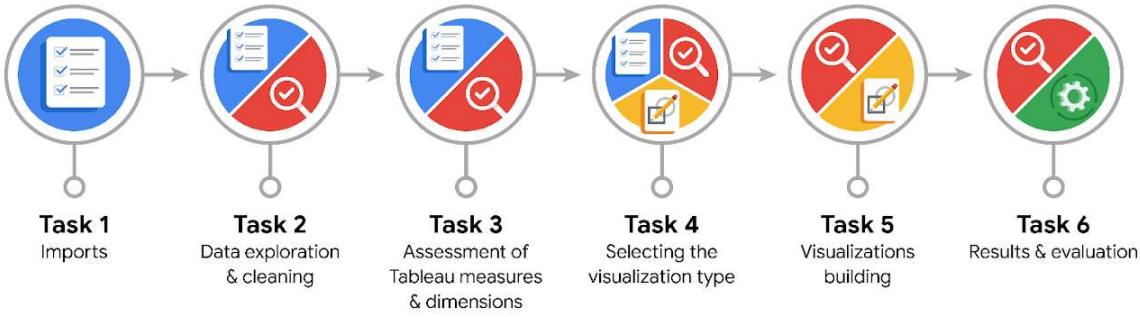
Relevant Interview Questions

Completing the end-of-course project will help you respond these types of questions that are often asked during the interview process:

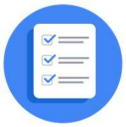
- How would you explain the difference between qualitative and quantitative data sources?
- Describe the difference between structured and unstructured data.
- Why is it important to do exploratory data analysis?
- How would you perform EDA on a given dataset?
- How do you create or alter a visualization based on different audiences?
- How do you avoid bias and ensure accessibility in a data visualization?
- How does 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: Plan Stage

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

The data columns and variables are a mixture of int64, float64, and object. Based on the preliminary investigation completed, it appears that the following variables are most relevant to the project deliverable: 'label' and the independent variables 'drives,' 'total_sessions,' 'driven_km_drives,' 'duration_minutes_drives,' 'activity_days,' and 'driving_days'.

- What units are your variables in?

Some variables are represented in units of kilometers, days, minutes, and the number of occurrences in the month.

- 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?

Initial investigation has found that there are 700 missing values in the 'label' column, so a strategy on how to handle them will be required, whether through imputation, dropping the rows if they have little impact on the overall dataset, etc. Additionally, looking at the data pertaining to the amount of driving the users complete, they appear very high, much higher than would be expected of the

average driver. Furthermore, there appear to be some significantly high values that may be indicative of outliers that require further investigation.

- Is there any missing or incomplete data?

The 'label' column is missing 700 rows of data. However, initial investigations indicated that this missing data is random and does not appear to have a significant impact on the overall quality of the dataset.

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

The variable types differ between int64, float64, and object; however, the formatting of the data is consistent.

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

Investigating and handling outliers, missing values, normalizing values if necessary, acquiring additional data if required, and engineering new variables out of existing ones will likely all be required during the EDA process.

PACE: Analyze Stage

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

First and foremost, organization and planning are key. Knowing what steps you'll need to complete to ensure the dataset is fit for analysis is crucial. Although having a systematic set of steps is important, it's unlikely that the steps will follow such a rigid order. Preparing for handling missing values, such as determining what type of missing values you're dealing with (missing at random, missing not at random, etc.), is important when devising a way to resolve it. Determining whether values need to be normalized for use in a machine learning model and inspecting and understanding outliers is also important when performing effective EDA.

- 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.?

Based on the outcome of the previous analysis, it was recommended to collect additional data from Waze users, ensuring it is representative of the general population of users at large. Additionally, inquiring about how the initial data was collected would be useful, as well as whether or not it pertains to a specific subset of Waze users, as the data appears to indicate. Furthermore, joining some of the engineered variables from the initial analysis might be beneficial. In terms of structuring, the variables in the data appear to be formatted correctly; however, the float64 variables could potentially be converted to int64 to reduce the size of the dataset, such as with the 'total_sessions' variable. The precision of a float64 for this variable is not required.

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

For a more technical audience, Histograms, Box Plots and Density Plots could be used to show information such as distributions, outliers, and metrics such as means, median and standard deviation. For the stakeholders, charts such as Line Charts, Bar Charts and Pie Charts could be useful as they're easier to understand.

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?

For this particular stage of the project, we will need to construct at least histograms and boxplots. Additionally, scatterplots and pie charts may be necessary, depending on specific requirements.

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

The data seems to be already formatted in a way that requires minimal effort before plotting. Typically, it's essential to verify that the data is free of errors, missing values, significant outliers, or normalized values. Importing the correct packages, such as seaborn and matplotlib, is crucial. Additionally, creating new variables may be necessary to plot certain visualizations if the existing variables in the dataset do not provide sufficient information on their own.

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

At this stage of the analysis, most continuous variables are relevant as their distribution and potential correlations with other variables are unknown. Additionally, the categorical variables "device" and "label" are applicable and can be visualized using a pie chart.

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

Impute the missing values with the mean, median, mode, or ±1.5*IQR (or another percentile). Delete the rows with missing data if there are very few or if it has been determined that they do not have a significant or tangible impact on the analysis. Leave the missing values as is if they do not interfere with the analysis and if there are not too many.

PACE: Execute Stage

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

Many continuous variables exhibit significant right-skew and are not normally distributed, indicating that much of the data is clustered on the left with prominent outliers on the right. In contrast, some variables follow a uniform distribution.

The "label" column contains missing values. Comparing the two groups (rows with missing values and those without), the statistics suggest that the differences between them are negligible. This may also indicate that the data is "missing at random" and not due to data entry errors.

The number of days since onboarding for users with 40% or more of their total sessions occurring in the last month is uniformly distributed, indicating that these long-time users used the app significantly in the last month. Why?

The data indicates that the population of users in this data drives a lot more than would be expected of the average driver. This raises questions about whether or not this data is biased toward certain groups, such as rideshare users or other independent contractors running long-distance commuting services.

The "sessions" and "activity_days" features do not exhibit the expected mirrored distributions.

Interestingly, retained users have fewer drives than churned users. It was discovered that distance driven per driving day appeared to positively correlate with churn; users who drove longer trips on their driving days were more likely to churn. However, the number of driving days had a negative correlation with churn; users who drove more in the last month were less likely to churn. This could possibly indicate that the length of time a user uses the app, or the length of the active session for those driving longer distances, could be indicative of an issue causing dissatisfaction.

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

Follow-up on why the usage of the app in the last month for tenured users was so high is necessary. What was different about this particular month compared to others? Investigate the data collection strategy used, such as random sampling, systematic sampling, clustering, stratified sampling, etc.

Examine a subset of users still using the app who have similar characteristics to those that churned, specifically those who drove more on their driving days, and investigate if they are currently experiencing any particular issues with the app. Since these individuals tend to churn more, this may provide insight into why.

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

It would be interesting to research a subgroup of active Waze app users with similar patterns to those that churn. This research may shed light on why they are churning, such as identifying a particular shortfall of the app when used for long periods of time. Perhaps for those driving the most on their driving days, the app is malfunctioning after a certain period of time or not providing them with what they require. Investigate why so many people suddenly used the app significantly more than usual last month. What was different about this month?

- How might you share these visualizations with different audiences?

For a more technical audience, histograms and boxplots are effective in showcasing the distributions and specifics of the variables, such as mean, median, mode, IQR, outliers, etc. For a more general audience, higher-level visualizations that focus on the broader context and overarching story would be a more effective approach.