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

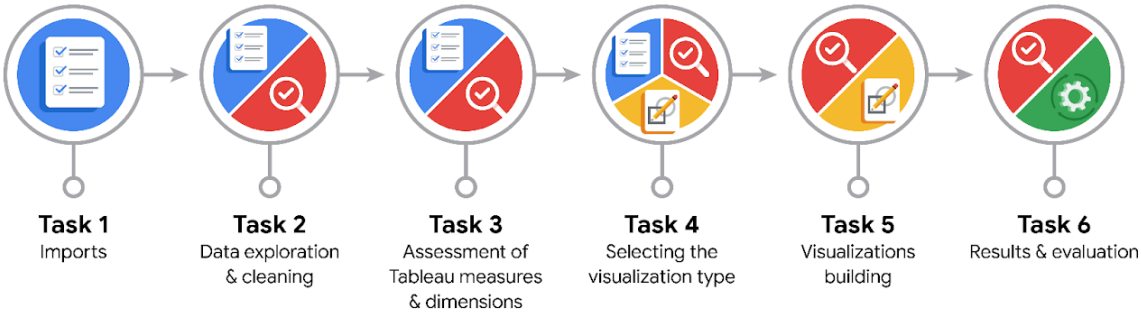
Relevant Interview Questions

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.



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 include: `ID`, `label`, `sessions`, `drives`, `total_sessions`, `n_days_after_onboarding`, `total_navigations_fav1`, `total_navigations_fav2`, `driven_km_drives`, `duration_minutes_drives`, `activity_days`, `driving_days`, and `device`.

For analyzing user churn, the most relevant variables are `label` (our target), `sessions`, `drives`, `total_sessions`, `n_days_after_onboarding`, `driven_km_drives`, `duration_minutes_drives`, `activity_days`, `driving_days`, and `device`. These variables provide insights into user behavior and app usage.

- What units are your variables in?

- `sessions`, `drives`, `total_sessions`, `n_days_after_onboarding`, `activity_days`, and `driving_days` are counts.
- `driven_km_drives` is in kilometers.
- `duration_minutes_drives` is in minutes.
- `ID` is a unique identifier.
- `label` and `device` are categorical.

- 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?

  - I presume that users with higher app usage (more sessions, drives, etc.) are less likely to churn.
  - I also suspect that longer-tenured users might exhibit different behavior patterns than newer users.
  - I expect that the distribution of many of the numeric columns will be right skewed.
  - I expect that there may be a difference in churn rate between IPhone and Android users.

- Is there any missing or incomplete data?

  Yes, the `label` column has missing values. This will need to be addressed.

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

  Yes, the dataset is in a structured CSV format, which is consistent.

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

  I will need to use summary statistics, data visualization (histograms, box plots, scatter plots, pie charts), and data cleaning techniques (handling missing data, outlier detection).

PACE: Analyze Stage

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

  - First, I'll clean the data by handling missing values and outliers.
  - Then, I'll explore the distribution of each variable using histograms and box plots.
  - Next, I'll examine relationships between variables using scatter plots and bar charts.
  - Finally, I will generate new features to better understand the data, such as km per driving day.

- 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.?

  - At this point, I don't think I need to join any additional datasets.
  - I will need to filter out invalid data, such as extremely high values for `km_per_driving_day`.
  - I will also need to create new columns, such as `km_per_driving_day`, and `percent_sessions_in_last_month`.

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

  - For the Director of Data Analysis, I'll use clear and concise visualizations that highlight key trends and insights.
  - Histograms and box plots will be useful for showing distributions and outliers.
  - Bar charts and pie charts will be effective for comparing categorical data.
  - Scatter plots will be used to show relationships between numeric variables.

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?

  I'll need to create histograms, box plots, scatter plots, pie charts, and bar charts.

  At this point, machine learning algorithms are not required.

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

  I'll use Python libraries like `matplotlib` and `seaborn` to create the visualizations.
  I'll ensure that the data is properly formatted and cleaned before plotting.
  I will utilize functions to reduce redundant code.

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

  `label`, `sessions`, `drives`, `total_sessions`, `n_days_after_onboarding`, `driven_km_drives`, `duration_minutes_drives`, `activity_days`, `driving_days`, and `device`.

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

  I will remove rows with missing `label` values, as the `label` is essential for our churn analysis.

PACE: Execute Stage

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

The overall churn rate is around 17%. Users who drive longer distances per driving day are more likely to churn. Users who use the app more frequently are less likely to churn. There are some data inconsistencies, such as the difference in maximum values for `driving_days` and `activity_days`. A large number of long time users had a high percentage of their sessions in the last month.

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

Investigate why long-distance drivers are more likely to churn and address any potential issues. Implement strategies to increase user engagement and encourage more frequent app usage. Review the data collection process to ensure consistency and accuracy.

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

• Why did so many long-time users suddenly increase their app usage in the last month? • What are the specific reasons for churn among long-distance drivers? • Are there any demographic or other user characteristics that correlate with churn?

- How might you share these visualizations with different audiences?

For the Director of Data Analysis, I'll provide a detailed report with key findings and recommendations, along with the visualizations. For other stakeholders, I'll create a more concise presentation with key takeaways and visual summaries. I will make sure to use clear and concise language in all presentations.