**Course Two**

# Get Started with Python



# 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 2 PACE strategy document
* Answer the questions in the Jupyter notebook project file
* Complete coding prep work on project’s Jupyter notebook
* Summarize the column Dtypes
* Communicate important findings in the form of an executive summary

# 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:

* Describe the steps you would take to clean and transform an unstructured data set.
* What specific things might you look for as part of your cleaning process?
* What are some of the outliers, anomalies, or unusual things you might look for in the data cleaning process that might impact analyses or ability to create insights?

**Reference Guide**

This project has three tasks; the visual below identifies how the stages of PACE are incorporated across those tasks.



**Data Project Questions & Considerations**

**PACE: Plan Stage**

* How can you best prepare to understand and organize the provided information?

Begin by exploring your dataset and consider reviewing the Data Dictionary.

1. **Understand the Business Context**: Clearly define the project's goal (churn prediction) and how churn is defined by Waze.
2. **Detailed Data Review**: Beyond basic head() and info(), thoroughly examine describe() output, check unique values for categorical columns, and understand variable relationships.
3. **Review and confirm the Data Dictionary**: Confirm the meaning, units, and potential range of values for every column. If anything is unclear or missing from the provided dictionary, that's when we'd add to it or seek clarification.

* What follow-along and self-review codebooks will help you perform this work?

1. **Python Libraries**: Core libraries like Pandas (for data manipulation), NumPy (for numerical ops), and Matplotlib/Seaborn (for visualization).
2. **Interactive Environments**: Use *Jupyter Notebooks* or *Google Colab*. They allow you to combine code, output, and explanations, making it easy to follow your own logic and review.
3. **Clear Documentation**: Use *Markdown cells* within your notebooks to explain steps, observations, and decisions. This acts as your personal "codebook."

* What are some additional activities a resourceful learner would perform before starting to code?
  + **Formulate Hypotheses** (optional): Brainstorm questions you want the data to answer (e.g., "Do users who drive fewer days churn more?").
  + **Sketch Out Analysis Plan**: Outline the high-level steps: data cleaning, initial exploration, feature engineering, modeling, and evaluation.
  + **Research Domain Knowledge** (optional): Learn more about typical Waze usage patterns, factors influencing app churn, and common metrics in mobile analytics.
  + **Consider Data Limitations**: Think about what questions the data cannot answer and what biases might exist.

**PACE: Analyze Stage**

* Will the available information be sufficient to achieve the goal based on your intuition and the analysis of the variables?

Yes, the dataset contains enough information to build a baseline churn prediction model. However, several integrity issues were identified, such as users with positive kilometers but zero recorded sessions, drives, or driving\_days. These anomalies must be addressed through cleaning or validation before modeling. While the available variables are sufficient for initial work, additional historical (multi-month) data would provide deeper behavioral context and improve model reliability.

* How would you build summary dataframe statistics and assess the min and max range of the data?

We would use the .describe() method in Pandas. The minimum and maximum values in its output directly show the range for each numerical variable. This allows quick detection of outliers and validation of whether values fall within logical limits.

* Do the averages of any of the data variables look unusual? Can you describe the interval data?

The averages themselves are not problematic, but they reveal skewness due to extreme outliers. For example, maximum values for sessions, drives, and driven\_km\_drives are far higher than their means, suggesting the presence of heavy users. In addition, some averages may be misleading if underlying variables contain contradictions *(e.g., users with kilometers logged but zero drives)*. The numerical variables in this dataset represent ratio data, not interval data. Ratio data has a meaningful zero point *(e.g., zero kilometers means no distance traveled)* and allows for meaningful comparisons using ratios *(e.g., 10 km is twice 5 km)*.

**PACE: Construct Stage**

**Note**: The Construct stage does not apply to this workflow. The PACE framework can be adapted to fit the specific requirements of any project.

**PACE: Execute Stage**

* Given your current knowledge of the data, what would you initially recommend to your manager to investigate further prior to performing exploratory data analysis?

We recommend first investigating the significant inconsistencies found in the data. Specifically, many users show a positive value for driven\_km\_drives but have **zero** drives or driving\_days. This is a critical data quality issue that must be addressed before any analysis can be trusted.  
In addition, the **700 missing values** in the label column also need to be investigated. Since label is our target variable for predicting churn, understanding why these values are missing is essential. They may represent new users who have not yet been classified, or they could be the result of data collection errors. Knowing the reason behind these missing values will determine whether these rows should be removed, imputed, or treated as a distinct group. Resolving both the inconsistencies and the missing labels will provide a stronger foundation for reliable analysis and modeling.

* What data initially presents as containing anomalies?

There are outliers in columns such as sessions, drives, total\_sessions, driven\_km\_drives, and duration\_minutes\_drives that could skew the analysis and model performance. In addition, some columns show inconsistencies. For example, a few users have zero drives or sessions but still show positive driving days or kilometers driven, while over 1,000 users have significant drives and kilometers but zero driving days. These are not just outliers, since these fields are expected to align, the mismatches suggest potential data errors that need to be validated and cleaned before modeling.

* What additional types of data could strengthen this dataset?

1. Having historical data of users, rather than just a single monthly snapshot, would likely improve the prediction accuracy of our machine learning model. It would let us see trends in user activity leading up to churn, like a gradual decrease in sessions or drives. This helps us understand the "why" much better than a single point in time.
2. Having data on app and feature usage would also be valuable. Beyond just monthly totals, more granular data—such as the frequency of app use within a day or week, or engagement with specific Waze features—could provide deeper understanding. For example, user interactions like actively reporting traffic, using the carpool feature, connecting with friends, or submitting bug reports or contacting customer support (a potential indicator of frustration leading to churn) can reveal important behavioral patterns.
3. Having user profile data—such as age, general location, or the types of routes they typically take (e.g., daily commutes vs. long-distance travel)—could also provide valuable context for understanding user behavior.