# Purpose

We would like to evaluate the technical and presentational skills of the candidates by emphasizing their ability to design and deliver a technical solution as an end-to-end pipeline in production.

# Deliverables

* A powerpoint presentation and a github link with your code.
* Any supplementary information (e.g. document) that is not included in the presentation but contains valuable information regarding the problem approach, way of thinking of the candidate etc.

# Presentation agenda

* 1 h 10 min interview, broken down into:
  + 5-minute introduction of the team and role
  + 20-minute presentation of technical task
  + 30-minute interviewer questions
  + 15-minute candidate questions

Candidates will be given a week to complete a technical task. Please remember there will be no bonuses for candidates who hand in their submissions early, so please take your time. If you have any questions or if something is unclear, please do not hesitate to contact us via email.

# Task Description

## Introduction

You are an ML engineer working for a company that uses **survey data** to understand/analyse customer behavior around luxurious Technology & Electronics brands. Your stakeholders (sales/marketing reps, product ops, analysts etc.) are not particularly tech savvy, meaning that their choice for delivering data to you is through csv/xlsx files. The various departments of the company are running as data silos. You are assigned your 10th clustering task for the month when suddenly an idea pops up into your head:

“To transform/automate this repetitive task into a data/ML pipeline by considering the different current or even future stakeholders and their specific needs.”

## Data description

To get a feel for how survey datasets look like, we have provided a sample which can be found in the accompanying file, [dataset](https://docs.google.com/spreadsheets/d/16CuGc4OVgAVwWKEencdiO6efCbr1ILqhDgBArlxzkug/edit?gid=534033959#gid=534033959)“dataset*.*csv”. There are a few things to note:

1. The dataset contains answers of a set of individuals in 2 surveys: a) “Core” b) “Core Recontact”, where “Core Recontact” is a recontact survey of the main “Core” dataset. By that terminology, we mean individuals who answer Core and are asked whether or not they would like to answer another survey, namely Recontact.
2. These individuals are otherwise referred to as *respondents*, who filled out a survey by answering a series of *questions*. For each question there are 1 or more different options that can be selected by a respondent and we call these options *datapoints*.
3. Each row in the dataset corresponds to a respondent.
4. The survey each question belongs to is denoted by its prefix, where “core” and “core\_re” denote Core & Core Recontact surveys respectively.
5. Within each survey, datapoints that belong to the same question share the same prefix e.g., datapoints q3\_1 and q3\_3 are answer options 1 and 3 for *question* q3 while *datapoint* q1\_1 represents the first option for question q1.
6. We say that a datapoint is *selected* by a respondent if its value in the corresponding column is set to 1.0. If a datapoint has not been selected, the answer is 0.0.
7. There are some constraints in the data (for the sake of simplicity, you do not have to worry about this for now):
   1. Constraints that depend on the nature of questions. For example, some questions are *single-selection* i.e., their options are mutually exclusive and thus a respondent must only select exactly one option. Others are *multiple-selection* questions (0, 1 or more options can be selected).
   2. Constraints that depend on the logical flow of questions in a questionnaire. These are referred to as “routing constraints”. For instance, when a respondent has answered that they do not own a car, this implies that in a follow-up question asking about the fuel they use, the response must logically be “no fuel” (e.g. “None of the above”).
8. When a respondent has not answered a question because they did not participate in the survey (you will notice that the response rates of “recontact” surveys are low), the value of the respective datapoint is NaN.

## Solution design

Consider a solution where you need to segment/cluster the respondents of the dataset.csv file into categories according to their purchase behavior/interests around luxurious Technology & Electronics brands. Consider each of the columns of the above dataset, a feature that is related to the purchase behavior/interests and assume that not all columns might be useful for your analysis.

Some ideas for your end-to-end pipeline:

* Ensure your solution is parametric and configurable. Imagine that this pipeline would need to run for other clustering scenarios too.
* Ensure your solution is production-ready with proper service decomposition. Avoid Jupyter notebooks and monolithic designs.
* Include unit/integration tests. You may use AI for that to speed up your work.
* Ensure proper deployment of your solution.

Some potential discussion points during the task presentation:

* Ingestion/storage layer: How to consume data and store it for further usage (database, s3 storage etc.).
* Processing layer: How to make the raw data ready to be processed by the ML service.
* Output/metrics/visualization layer: How you would want your output to be presented to your fictional stakeholders.
* Any deployment details and/or performance considerations are more than welcome to be discussed/addressed here.

Feel free to be creative when you present the segments/clusters story and pick information from any Audio&Sound/Photography/Mobile/Computing/Wearables/Home Technology brands. You may use an AI assistant to produce some fake insights for each of the produced clusters. Do not focus so much on presenting different models, or a perfect cluster separation but be prepared to justify your final model selection. Instead focus on your data quality. Check various scenarios, from omitting incomplete rows/columns to find a way to synthesize the missing piece of information (imputation techniques, synthetic data generation). You may also use an AI assistant for data generation but be prepared to be asked questions on how you guided it to fill in the gaps.

# Code of Conduct

We expect candidates to complete this assignment individually and without help from others. Candidates are free to consult online sources of information, however we will assume that candidates have a full understanding of any and all parts of their submission.

Best of luck with your assignment and we are looking forward to hearing from you!