# Context

You start working in the sales team of a solar company as **the Data Lead** and one of your first assignments is to build the basic data architecture of the department, to start obtaining insights from the data available.

The main objective of the company is to **sell and install solar panels to domestic customers.** The company covers most of the provinces of the country, although there is a concentration of presence in the metropolitan areas of Barcelona and Madrid.

The solar company has a very simple product portfolio. Currently, they have one unique option of solar panels type. However, the installation is personalized to each roof, so the number of panels included can vary from one installation to another. In addition, the installation cost also can depend on the installation complexity, as more installation hours requires a higher cost.

The customers also can pay for the installation by cash covering the whole import before the installation is done, or they can finance the installation, which the company finances at a 5% interest rate.

Focusing solely on the sales process after a lead is acquired, two teams are involved: the sales team, responsible for increasing the lead-to-sale conversion, and the installer team, responsible for completing physical installations. Installations may be carried out by either an internal team or an external team, as the internal team's coverage is limited. While internal team installations can mean a lower cost and a highest customer satisfaction, external teams provide the flexibility to install in peak moments.

Each installation is related to:

- **The peak power**: the maximum energy amount that the installation can produce in a single unit of time.
- **The installation cost.**
- **The number of panels:** the number of panels of an installation varies according to the modules that fit on the roof. Normally, as more modules are included, more energy can be produced, although this can vary depending on the roof orientation and climate zone.

On the other hand, it is considered a sale when the project is validated by the technical team. This means that the installation team receives the order to perform the installation. If we look closely at the sales process, the phases involved are:

1. An offer is sent to the customer with the number of panels, configuration, payback and installation price.
2. If the customer likes the offer a contract is sent to the customer to perform the purchase.
3. In the event of a financed installation, a second contract, corresponding to the financed part, is sent to the customer, normally on the same day as the first one.
4. The customer can sign the contracts to perform the purchase.
5. A sales visit can be made to the customer, to close the deal and fine-tune the final aspects of the solar configuration. This can be done before or after the contract's signature.
6. After the previous steps, a technical internal team reviews the project and creates the documentation for the installer team.
7. If everything is OK the project is validated and sent to the installer team.
8. In any moment of the funnel a lead can leave the sale process. In that case, for some of them the company stores the KO reasons. Also, if the lead changes his/her mind after purchasing the product (in the following 15 days) and does not want to purchase the product, the company can cancel the sale. In this case, we talk about sale dismissal.

As you work in the sales team you have the objective of helping the team with ideas and prioritization **to increase the lead to sale conversion rate** and track the metrics to know the status quo.

Therefore, you have access to a dataset containing data about the sales phases, the KO reasons and the product specifications in order to exploit it.

# Dataset

## Sales phases table

The table contains all the leads and sales that the company has acquired during 2024. For each one of the leads the table contains the dates that the lead enters in each phase, the KO date (if applies) and some attributes related to the lead.

The fields are:

- **lead_id**: a unique numeric identifier of each lead.
- **financing_type**: if the offer payment type is cash or is financed.
- **current_phase**: the current (or last phase) of the lead.
- **phase_pre_ko**: the phase before the KO. In the case of non KO leads, this is the current phase.
- **is_modified:** if the initial offer was modified at some point (e.g. the number of panels was changed).
- **offer_sent_date:** the date when an offer was sent to the customer.
- **contract_1_dispatch_date:** the date when the first contract (related to the installation) was sent to the customer.
- **contract_2_dispatch_date:** the date when the second contract (in case of financed sales) is sent to the customer. This matches the same date as the contract 1 or not.
- **contract_1_signature_date:** the date when the first contract was signed by the lead.
- **contract_2_signature_date:** the date when the second contract was signed by the lead.
- **visit_date:** the date when the sales expert went to the lead address in order to close the deal or perform an evaluation.
- **technical_review_date:** the date when the technical team starts reviewing the project.
- **project_validation_date:** the date when the project is validated and therefore considered a sale.
- **sale_dismissal_date:** the date, if applicable, when the customer has cancelled the purchase.
- **ko_date:** the date that the customer leaves the sales funnel. In case of KO leads only.
- **zipcode:** the zip code of the customer.
- **visiting company:** if the sales expert is internal or external.
- **installation_peak_power_kw:** the peak power of the installation. The maximum power that the installation can produce (with the maximum

solar radiation). In addition, the higher the power the higher the energy that the installation will produce (and higher the savings).

- **installation_price:** the price of the installation that the lead must pay.
- **n_panels:** the number of panels that has the installation.
- **ko_reason:** the reason for not performing the sale successfully.

In addition, you will also find two extra tables that can help you with your analysis:

- **A weather data table**: consists of a list of different weather characteristics per day and zip code for 2024.
- **A zip code table:** consists of a table that relates each zip code with the location, the province and the autonomous community.

# Objective

The objective of this assignment is to **design, build and use** the data model to obtain insights from the solar panels sales business process.

Therefore, as the Data Lead in the area you will need to:

- **Define the data model that the business process will have in the Data Warehouse.** As the main table (*the sales phases table*) is obtained in the flat schema form you will have to design which is the best configuration for analytics.
- **Create the data pipeline to transform the input data to the final configuration.** You will need to create the pipeline that ingests the data and performs the necessary operations to have in the output database the model that you design. You can choose the methodology that you prefer: classic ETL data pipelines or a DBT configuration.
- **Implement data cleaning and data quality checks in the ingestion pipeline** to avoid having low quality data in the Data Warehouse. As a data modeler, you will need to investigate if it makes sense to filter rows or fill nulls in the pipeline, and then apply these rules in the ingestion and transformation scripts.
- **Optimize the Data Warehouse** including indexes and primary keys. To have a high performant tool.
- **Use the data model that you have created in your MySQL database to answer business questions:**
    - Which are the top 5 KO reasons, and which percentage of total KO represent each one?
    - Which has been the month with more sales of customers that have contracted the cash financing type product? How many sales did that month have?
    - To know if the price of installation affects the financing type, can you obtain a table that shows for each financing type the average installation price? Only consider the individual households for this indicator.
    - Only considering the zip codes with more than 5 leads, which is the average peak power, average installation price, number of leads per province?

- Is average temperature correlated to the sales conversion? Provide a table that contains 3 columns: zip code, sales conversion and average temperature. The sales conversion is calculated dividing the sales done / leads of that zip code. To get the sales done you can filter the leads that are in the “Validated project” phase.

## Deliverables

You must deliver:

- **A pdf document explaining the data model that you have designed**. Explain the benefits and the reason behind it and illustrate the schema using ChartDB or an equivalent tool. List the primary and foreign keys of your design.
- **Python or SQL scripts that ingest the input data as csv and write the final data model in a database**. You can choose to do the transformations in python or to load the raw data in the database and do the transformations in SQL (DBT). The scripts must follow the best practices and a naming convention. Explain what the scripts do inside the same files as comments. If you create functions comment on them explaining their objective, input variables and output objects. Inside these scripts include 2-3 cleaning / data quality operations.
- **A pdf document explaining which cleaning operations** you have decided to include in the data pipeline and which benefits do they provide.
- **A pdf explaining which indexes** you have decided to include in the Data Warehouse and why.
- **A file containing the SQLs to obtain the 5 indicators** explained in the *Objective* section. You must provide the SQLs and a screenshot of the query result. If the result has many rows, it is enough to show the first ones in the screenshot (e.g. the 4$^{th}$ bullet point).

*The pdf bullet points you can deliver them in a single file.

## Grading (over 10)

- Data model design. **(1 points)**
- ETL/ELT code. **(2 points)**
- Cleaning operations**. (2 points)**
- Database optimization - indexes**. (1 point)**
- Indicator SQLs. **(3 points)**