1. **Business understanding & problem framing:** what is the context of this problem and why are they trying to solve it?
2. **Exploratory data analysis & data cleaning:**what data are we working with, what does it look like and how can we make it better?
3. **Feature engineering:**can we enrich this dataset using our own expertise or third party information?
4. **Modeling and evaluation:**can we use this dataset to accurately make predictions? If so, are they reliable?
5. **Insights & Recommendations:**how we can communicate the value of these predictions by explaining them in a way that matters to the business?

Here are the key takeaways from the meeting:

* Your client is **PowerCo**- a major gas and electricity utility that supplies to small and medium sized enterprises.
* The energy market has had a lot of change in recent years and there are more options than ever for customers to choose from.
* PowerCo are concerned about their customers leaving for better offers from other energy providers. When a customer leaves to use another service provider, this is called **churn**.
* This is becoming a big issue for PowerCo and they have engaged BCG to help diagnose the reason why their customers are churning.

**Task 1**

* Determine the client data needed for analysis
* Outline the techniques you'll use to investigate your client's problem
* Write an email to your Associate Director summarizing your approach

First things first, you and Estelle need to understand the problem that PowerCo is facing at a deeper level and plan how you’ll tackle it. If you recall the 5 steps in the Data Science methodology, this is called “business understanding & problem framing”.

Your AD wants you and Estelle to email him by COB today outlining:

1. the data that we’ll need from the client, and
2. the techniques we’ll use to investigate the issue.

You must formulate PowerCo’s issue as a problem using the 5 step data science process and lay out the major steps needed to test it.

1. What do you think are the key reasons for a customer deciding to stay with or switch energy providers? For example: price, is it clean energy, customer service, location etc.
2. What data do you think would be useful in order to investigate these key reasons? E.g. customer purchasing trends over past 5 years, location of business etc.
3. If you were to get this data, how could you analyse or visualize it to test whether these reasons may have an impact on churn?

In order to test the hypothesis of whether churn is driven by the customers’ price sensitivity, we would need to model churn probabilities of customers, and derive the effect of prices on churn rates.

We would need the following data to be able to build the models.

1. Customer data - which should include characteristics of each client, for example, industry, historical electricity consumption, date joined as customer etc.
2. Churn data - which should indicate if customer has churned
3. Historical price data – which should indicate the prices the client charges to each customer for both electricity and gas at granular time intervals

Once we have the data, the work plan would be:

1. We need to define what price sensitivity is and calculate it
2. We need to prepare the data and engineer features
3. Then, we can test our hypothesis using a binary classification model (e.g. Logistic Regression, Random Forest, Gradient Boosted Machines to name a few)
4. We would choose a model from one of the tested algorithms based on the model complexity, the explainability, and the accuracy of the models.
5. With the trained model, we would be able to extrapolate the extent to which price sensitivity influences churn

**Task 2**

* Use python to analyze client data
* Create data visualizations to help you interpret key trends

# **Your AD is giving you more responsibility!**

Well done for your initial understanding of the case with PowerCo. After reviewing your project plan, the AD would like you lead on the Data Science deliverables for the rest of the project.

The AD would like you to investigate whether price sensitivity is the most influential factor for a customer churning, and if not, to what extent does price sensitivity influence churn.

Before we begin on this task, what exactly is price sensitivity?

**Explanation**

Getting set up - This task is focused on exploratory data analysis of the client and price data provided:

Analysis - Once you’ve run the cells provided, it was your job to build on this exploratory analysis:

* The visualization provided by Estelle shows how many companies churned vs. how many companies did not churn. We can see from this that the churn rate is approximately 10%. This is actually a very good churn rate, the closer the rate is to 0%, the better.
* The next series of visualizations were created in an attempt to try and dive deeper into how churn changes based on other factors (using other columns). This is useful for us to investigate because it may help us to understand factors that drive churn.
* In the notebook we visualize churn vs. sales channel, contract type, number of products, number of years and origin/contract offer.
* For example:
  + We see that for sales channel, there are some sales channels that yield customers churning but there are also other sales channels that have no customers churning.
  + For contract type, we see quite an even split for customers churning. This is interesting because this may suggest that contract type is not a driving factor towards churn rate.
* Additionally, for some columns their distributions with churn rate included. This is useful for us to understand because based on the distribution of a column, this could affect our feature engineering later.
* We look at the distribution of consumption, subscribed power and forecast in the notebook.
* For example:
  + We notice that the distribution of consumption is very skewed, this is called a positive skew since it is biased towards lower values on the x axis.
  + This is interesting because you may decide to treat this column to reduce the skewness later on during feature engineering. But also because we may want to visualize if there are any outliers within this column.
  + To investigate outliers, we use a boxplot. From the boxplot we can see that with the column as it is there are definitely some outliers. Once again this is interesting because we may choose to remove some of these outliers later.

# **Task 3**

# **Explanation:**

Set up:

* This task is focused on feature engineering, Estelle has provided a CSV dataset for you to use as a base for this task.
* You should download the notebook and CSV and start by running the cells within the notebook.
* Estelle has also provided some insight into a feature that would be interesting to add to the data. By running the cells in the notebook, this will create the feature that she described for you and provide a foundation for you to begin your own feature engineering.

Here is some context around the additional features that have been engineered in the notebook, to help you in the future:

* Firstly we have the average price changes across periods. This is a measure of the average price change by company between peak, mid-peak and off peak periods.
* We then take this idea one step further by creating another similar feature but instead of looking at the average price difference, we look at the maximum price difference across periods and months. This gives another way to look at the price changes across months.
* The reason why these 2 features could be useful is because they are another way of representing the variance of prices throughout the year. Imagine, if your utilities bill massively increased over winter, as a consumer you’d be annoyed and want to find a better deal!
* After this we continue feature engineering with some more concepts, including transformation of columns.
* To make predictions with a statistical or machine learning algorithm, all of the data must be converted to numeric data types.
* Therefore, we convert date into months and remove the raw date column, as we cannot use it in its original form.
* We also convert boolean columns into binary values.
* And we convert categorical columns into dummy variables. A dummy variable is a binary flag that indicates when a row matches the value from the categorical column that it was created from.
* As we saw during exploratory data analysis, the distribution of some columns was skewed. This is important to identify because when modeling data for prediction, based on the technique or algorithm that we use, there are sometimes assumptions within the data that we should follow.
  + One common assumption is that the columns within the data are normally distributed. Hence, if we find that columns are not normally distributed, we should treat these columns to try and transform them into a distribution that is more normal.
* Therefore, the next thing we do is transform some columns to have a closer to normal distribution. We do this using the logarithm function. As you can see from the visualisations, the newly transformed columns are much closer to a normal distribution than what they were earlier.
* Finally, we plot correlations of all the columns to see if we can identify any columns to remove. Columns that have very high correlations indicate an area to look out for. In this case, you may want to remove one of the columns, since they are likely both holding very similar information.

**Task 4**

This final task is focused on building the predictive model using the CSV file that Estelle has shared.

* This CSV file contains a set of cleaned and engineered features so that you can focus purely on training your predictive model.
* You should download the Jupyter notebook and CSV file and run the cells provided in the notebook.
* These cells will load the data and create train and test samples of the data.
* It is important to split your data into train and test samples so then you can measure how well the trained model performs on an unseen set of data.
* This is a massively important thing to do when building a predictive model, otherwise you will have no way of measuring how well your model is able to predict churn for new customers!
* The code in the notebook provides you with skeleton code to create the random forest classifier, but it is your job to fill in the details of the code by using the documentation site provided.
* By adding in values for parameters within the random forest and by fitting the model on the training data, you will have a trained model to predict churn!

Now the most important part, evaluation of the model:

* It is left for you to decide how to evaluate the performance of the model. In general, you want to use metrics that reflect honestly how well the model has performed.
* In the notebook we use 3 metrics, accuracy, precision and recall.
* The reason why we are using these three metrics is because a simple accuracy measure (what percentage did I predict correctly) is not always a good measure to use.
* To give an example, let's say you're predicting heart failures with patients in a hospital and there were 100 patients out of 1000 that did have a heart failure.
* If you predicted 80 out of 100 (80%) of the patients that did have a heart failure correctly, you might think that you've done well! However, this also means that you predicted 20 wrong and what may the implications of predicting these remaining 20 patients wrong? Maybe they miss out on getting vital treatment to save their lives.
* As well as this, what about the impact of predicting negative cases as positive (people not having heart failure being predicted that they did), maybe a high number of false positives means that resources get used up on the wrong people and a lot of time is wasted when they could have been helping the real heart failure sufferers.
* This is just an example, but it illustrates why other performance metrics are necessary such as precision and recall, which are good measures to use in a classification scenario like this.
* After calculating the 3 metrics, we can see that we’re able to accurately identify clients that do not churn, but not so accurately identify clients that will churn. Our model is predicting a high percentage of clients to not churn, when in fact they did!
* This tells me that the current set of columns are not a good set of features to predict churn. As the data scientist, it would normally be my job to go back and try to engineer a set of features that is able to predict churn more accurately.

Finally, we produce a feature importance chart to visualise which features were indeed useful within the model and which ones weren’t.

* We can see that net margin and consumption over 12 months were important, to name a few.
* However the price sensitivity features are scattered around and do not shine through as a main driver for churn in their current form.