# Goal

Bayer wants to increase the market share in Germany for its most important anti-allergy product, Snaffleflax®. A critical step in achieving this goal is understanding, measuring and predicting this market share against the three primary competitor products in the market.

Your task is to build a dataset that can be used to train a machine learning model. The machine learning model, which will be developed by your data scientist colleague, is a regression that predicts market share for Snaffleflax®. There are at least two types of signal the ML model can learn from to predict future market share: historical market share and historical sales and marketing activity, ie receiving email newsletters or face-to-face visits with sales reps. The historical sales and marketing activity is stored in a Customer Relationship Management (CRM) database.

# Data description

You are provided with several sets of raw data that needs to be processed into a single flat file that could be written to a CSV.

## Sales data

The sales data has the following schema:

|  |  |  |
| --- | --- | --- |
| **Field name** | **Type** | **Description** |
| acct\_id | string | The customer account UUID |
| product\_name | string | The product name as recorded by the distributor. Can denote any of the four products on the market |
| date | date | The date of the sale (rounded to the month) |
| unit\_sales | integer | the units sold for that account in the month |
| created\_at | timestamp | the creation timestamp of the record in the database |

This data is extracted every day from a REST API and saved in a file in the data lake that has a filename with the pattern <yyyy-mm-01>.json. Because of complexities in the distribution network, it frequently occurs that a sale that initially gets reported to a given customer is later voided, but this sales data is always “closed” at the end of the month, after which the data for the month does not change.

## CRM data

The CRM data has the following schema:

|  |  |  |
| --- | --- | --- |
| **Field name** | **Type** | **Description** |
| acct\_id | string | The customer account UUID |
| event\_type | string | The event type: f2f, workplace event, or group call |
| date | date | The date of the event |

This data is extracted every day from the CRM database. The whole table is extracted to a file called crm\_data.csv in the data lake and overwrites the previous version.

# Exercises

1. Write a Python script or Jupyter Python notebook to process the data into a CSV suitable for training the ML model described above. Be sure your output includes columns for the following:
   1. Market share (the target variable for the regression; should be in the range [0,1]). Use the simplifying assumption that the data provided covers the entire market.
   2. The lagged X-month average of market share, where X is a parameterized integer.
   3. The lagged X-month sum of events, where X is a parameterized integer.
   4. The lagged X-month weighted sum of events, where X is a parameterized integer and the weights are a list of numbers with length X.
2. How could you enhance the dataset with publicly available data? Describe in one or two sentences, but do not implement.

# Presentation

You will have 40 minutes to present your solution in a code walkthrough.

# Advice

You are welcome to ask questions via email, but in order to keep things fair, we may not be able to offer precise answers. Therefore, if you feel the problem statement or data is ambiguous, please try to make reasonable assumptions and describe them in your presentation.