# Data Science Case Study

## Overview

The objective of the IFF Data Science Case Study is to help you understand the type of problems we deal with on daily basis and allow you to showcase your capabilities. As one of the largest companies in the taste, scent, nutrition and specialty ingredient production industry, a lot of our data science projects deal with manufacturing and process data, trying to understand how we can improve processes, increase throughputs and productivity, or optimize production schedules. Process data, such as process parameter setpoints, sensor readings, and product quality testing results, are typically provided in the form of time series. Often our problems consist of understanding how these time series relate to each other. For example, how do process setpoints impact product quality?

Although this case study does not deal specifically with process data, it does involve time series data analysis, posing a very frequent generic question: how do two time series relate to each other?

## Instructions

- You may not discuss this case study with others or share it online.
- We strongly recommend completing work for this case study in Python, since it is the primary language used by the IFF Data Science team.
- Usage of open-source packages, online or book resources is allowed and encouraged! We are interested in your analytical approaches rather than memorization of programming syntax.
- Please return your output to this challenge within 5 days as a Python script or Jupyter notebook that can be executed by one of our Data Scientists to verify the answers you provide.

## Case Study

For improved financial planning, including purchasing and product pricing, it is important for us to understand and forecast our raw input ingredient costs. A lot of raw ingredients used in our manufacturing processes are petrochemically derived, i.e. they are derived from crude oil.

In this case study you are given prices for crude oil (“oil_prices.csv”) and prices for one of the petrochemically derived ingredients we frequently use in our manufacturing processes (“ingredient_prices.csv”). **We ask you to help us understand the correlation (and lag in the correlation) between the oil and the ingredient prices**.

You are welcome to provide any type of analysis you feel is appropriate, but below we have a few suggested questions you can address in your analysis. Please document and be prepared to talk about the topics you decide to address in your analysis.

- What is the quality of our data?
- What are the time series properties, such as stationarity, seasonality, auto- and cross-correlation?
- Does the time series behavior change over time?
- Can we forecast price of raw ingredient using price of oil?

## Evaluation Criteria

As typical of most data science projects, this case study poses a very open-ended question for which there is no "perfect" answer or approach. However, we want this exercise to give us a view into the way you address a problem using data, and there are some general traits that we are looking for:

- What is your thought process in understanding the data you are working with? (exploratory data analysis)
- How do you derive/validate/quantify conclusions from the data? (more formal statistical or model-based analysis)
- Do you deliver your conclusions in concise and simple language easily understandable by non-statistics business managers?
- Are you providing appropriate visualizations to support your observations and conclusions?
- Do you have any actionable recommendations based on your analysis with proposed next steps?
- Are you writing production worthy, modular (think functions and classes), readable (variable names, comments), and reproducible code?