Yieldigo

Dear prospective colleague,

we have created this test to get aligned with you and your data skills and talent. We expect you to spend less than 3 hours with it and find at least 8 errors in the data analysis part.

We realize we ask for quite a lot of your time, so we offer you a thing in exchange, too. After you finish the test, you'll have a chance to speak with one of our engineers, who'll provide you with some feedback, so maybe you gain more insight into whether a data position like this is your way to go.

1. Data analysis assignment

Raw data usually contain multiple data issues. Your goal, as a data analyst or engineer, is to find, discuss and ultimately resolve these issues before you start automatically uploading them into the system. As a rule, we require clients to have a complete sales history in 2-3 years, so checking the data in Excel is usually impossible - we typically upload the data to temporary tables in the database or load them as Pandas dataframes, so we can run complex and scalable analysis.

The most common types of errors we encounter are **negative margins**, **duplicate sales**, **different purchase prices within the same stores**, and **incorrect formatting of values**. However, we also encounter errors that you have no idea exist, and you have to look for them creatively.

Sample sales are based on real store data and are a cutout of the annual aggregation of 15 products that were sold at 5 stores within two different groups.

All necessary data with their specification can be found in the enclosed CSVs below.

Can you tell what errors do the files contain? Send us the resulting data analysis in the form of a PDF or data file as well as the underlying scripts. You can use any tool of your choice for the analysis, but bear in mind we are looking for a solution that is scalable and automatable down the road, so Excel is probably not a great idea. Also, we want to stress that the important aspect of the analysis is a discussion and description of your process. **We are looking for a critical thinker**, not someone who has only read some tutorials and just throws random functions on the problem.

Files description

zones.csv

The stores are divided into groups (internally called zones), over which price optimization then takes place. A different pricing policy may apply to the Hypermarket than to the Supermarket. Shops in one zone must have the same prices.

sites.csv

Stores with the same price over one product are sorted into price groups according to the ZoneId key. The shops are most often sorted by region and type of shop.

articles.csv

List of products sold that are linked to sales data via ArticleId.

sales.csv

Sales are the largest and most important table. Based on the contained data, subsequent price optimization is performed using mathematical models. Most often, the client stores sales in daily aggregations based on primary key: product number, day of sale, site, cost price, and sales price. Within these aggregations, of course, most errors occur and for the proper functioning of the mathematical model, it is necessary to eliminate these errors. The selling price "Price" and the cost price "CostPrice" are in this case given for 1 piece.

2. Scripting assignment

Code below is a Python implementation of one famous algorithm. Could you briefly walk us through its steps and tell us what the result would be?

```
def bb(arr):
    n = len(arr)
    for i in range(n-1):
        for j in range(0, n-i-1):
            if arr[j] > arr[j+1] :
                arr[j], arr[j+1] = arr[j+1], arr[j]
```